

Circadian input kinases and their homologs in cyanobacteria: Evolutionary constraints vs architectural diversification

Ivan Baca¹, Daniel Sprockett², and Volodymyr Dvornyk^{3§}

¹ Institute of Genetics, National Academy of Sciences of Moldova, Chisinau, Moldova,

²Department of Biological Sciences, Kent State University, Kent, OH 44242-0001, USA,

³School of Biological Sciences, University of Hong Kong, Pokfulam Rd., Hong Kong SAR, P.R. China

[§]Corresponding author

Email addresses:

IB: bidmd@yahoo.com

DS: dsprocke@kent.edu

VD: dvornyk@hku.hk

Abstract

The *cikA* gene encodes a protein relaying environmental signals to the central circadian oscillator in cyanobacteria. The CikA protein has a variable architecture and usually consists of four tandemly arrayed domains: GAF, histidine kinase (HisKA), histidine kinase-like ATPase (HATPase_c), and a pseudo-receiver (REC). Among them, HisKA and HATPase_c are the least polymorphic, and REC is not present in heterocystic filamentous cyanobacteria. CikA contains several conserved motifs that are likely important for circadian function. There are at least three types of circadian systems, each of which possess a different set of circadian genes. The originally described circadian system (*kaiABC* system) possesses both *cikA* and *kaiA*, while the others lack either only *cikA* (*kaiABC*^Δ) or both (*kaiBC*). The results we obtained allowed us to approximate the time of the *cikA* origin to be about 2600-2200 MYA and the time of its loss in the species with the *kaiABC*^Δ or *kaiBC* system between 1100 and 600 MYA. Circadian specialization of CikA, as opposed to its non-circadian homologs, is a result of several factors, including the unique conserved domain architecture and high evolutionary constraints of some domains and regions, which were previously identified as critical for the circadian function of the gene.

Running title: Circadian input kinases of cyanobacteria

Key words: CikA, prokaryotes, GAF, divergence, constraints

Introduction

Cyanobacteria are the simplest organisms known to have an endogenous circadian clock (Kondo and Ishiura, 1999). The model species *Synechococcus elongatus* PCC 7942 contains a cluster of three tandemly arrayed genes, *kaiA*, *kaiB*, and *kaiC*, which has been identified as a key element of the circadian system (Ishiura et al., 1998). Possession of a circadian clock has been shown to enhance the adaptive fitness of cyanobacteria in a wide variety of environmental conditions (Woelfle et al., 2004).

In addition to the three *kai* genes, which were thought to be indispensable for circadian oscillation (Ishiura et al., 1998; Kitayama et al., 2003; Xu et al., 2003), several other genes have been identified to control the input and output of the cyanobacterial clock (see Golden and Canales, 2003, for review). An evolutionary analysis of various components of the circadian system (Dvornyk et al., 2003; Dvornyk et al., 2004; Dvornyk and Knudsen, 2005; Dvornyk, 2006a) suggested that cyanobacteria have at least two types of the system, those with and those without *kaiA* (hereafter referred to as *kaiABC* and *kaiBC* system, respectively). The species lacking *kaiA* possess a timing mechanism, although it is less robust than the original *kaiABC* system (Holtzendorff et al., 2008; Axmann et al., 2009).

The *cikA* (circadian input kinase) gene encodes a bacteriophytochrome-like histidine kinase involved in the input signaling of the clock (Schmitz et al., 2000). CikA was reported to have three distinct domains: GAF, histidine protein kinase, and a receiver domain (Mutsuda et al., 2003). This structure is typical for bacteriophytochromes (Fankhauser, 2001). However, CikA is missing a conserved cysteine residue, which serves as a bilin ligand in the sensor domain of typical phytochromes. Because of this deletion, it is categorized as an unusual

bacteriophytochrome (Schmitz et al., 2000). A recent study showed that CikA senses light not by a chromophore binding to the GAF domain, but through detecting quinones (Ivleva et al., 2006). The concentration and redox state of quinones in a cell is light dependent. Mutants that are *cikA*-deficient have a shorter circadian period of gene expression and altered phasing of rhythmicity (Schmitz et al., 2000).

In this study, we have analyzed the occurrence, domain architecture, level of variation and phylogeny of the *cikA* gene in order to reconstruct the evolutionary history and to determine the evolutionary factors that have been operating on this component of the cyanobacterial circadian system. We have also attempted to estimate a timeline for key events in the evolution of both *cikA* and the entire circadian system. The present work provides new data about the probable functional importance of various structural motifs of the CikA protein, and significantly updates our knowledge about the evolution of the cyanobacterial circadian system as a whole.

Methods

DNA and protein sequences

The homologous sequences of the CikA proteins, 16S rRNA and 23S rRNA genes were retrieved from the GenBank non-redundant database using gapped PSI-BLAST (with 3 iterations) and BLASTN tools (Altschul et al., 1990; Altschul et al., 1997). The following GenBank accession numbers of the sequences from *S. elongatus* PCC 7942 were used as queries: AAF82192.1, AF132930.1 and CP000100.1, respectively. Only the sequences from completely sequenced cyanobacterial genomes were utilized for the phylogenetic analysis. Since CikA is a member of a large family of bacteriophytochromes, two criteria were used to filter the sequences for the subsequent analyses. First, the proteins should have at least three (GAF-HisKA-HATPase_c) of the four domains arrayed in the same order as in the originally

described CikA. Second, all these domains should display sufficiently high homology to *bona fide* CikA (bit score of 177 was used as a lower limit of homology). With such an approach, some proteins having formally higher similarity score (usually limited to the HisKA-HATPase_c domains) but lacking the above domain architecture were excluded from the analyses.

The sequences were aligned using MUSCLE (Edgar, 2004). The aligned 16S rRNA and 23S rRNA sequences were trimmed and concatenated. The CikA protein sequences were manually adjusted based on the available data about the protein's structure (Mutsuda et al., 2003) to match the putative domains. The *cikA* nucleotide sequences were aligned against the aligned protein homologs using RevTrans v. 1.4 (Wernersson and Pedersen, 2003) available online at <http://www.cbs.dtu.dk/services/RevTrans/>. The list of the used sequences is given in Supplementary data online Tables S1-S3.

Analysis of variation and phylogenetic reconstruction

The DNA substitution model that fitted the data best was determined for the concatenated 16S rRNA and 23S rRNA genes using the hierarchical test as implemented in the ModelTest 3.0 software (Posada and Crandall, 1998) and HYPHY (Kosakovsky-Pond et al., 2005). Based on the results of this test, the Tamura-Nei model of substitutions with gamma distribution (Tamura and Nei, 1993) and $\alpha = 0.40$ was used for further phylogenetic analysis of the concatenated rRNA genes. For the CikA protein homologs' phylogenetic reconstruction, two empirical amino acid replacement matrices were tested: WAG (Whelan and Goldman, 2001) and LG (Le and Gascuel, 2008). The latter yielded a tree with significantly better likelihood scores. In the reconstruction of the species tree, the 16S rRNA and 23S rRNA genes of the proteobacterium *Rhodobacter sphaeroides* ACC 17029 were used as an outgroup.

The rate of nonsynonymous nucleotide substitutions per nonsynonymous site (d_N) was calculated using the Pamilo-Bianchi-Li method (Pamilo and Bianchi, 1993; Li, 1993). The rate of synonymous nucleotide substitutions could not be estimated due to saturation. The MEGA 4 software package (Tamura et al., 2007) was used for the computations of d_N .

The phylogenetic tree of the Cika-like proteins (in total 1113 sites) was constructed using the maximum-likelihood (ML) algorithm implemented in the PHYML 3.0 software (Guindon and Gascuel, 2003). The phylogeny of the concatenated rRNA (4529 sites) genes was reconstructed using two approaches: the ML method as described above and the Bayesian relaxed clock phylogeny as implemented in the BEAST software (Drummond and Rambaut, 2007) with MCMC run for 10 million generations and trees sampled every 1000 steps. The reliability of tree topologies inferred with the ML approach was statistically evaluated using nonparametric bootstrap (100 replications) and the approximate likelihood-ratio test (aLRT) (Anisimova and Gascuel, 2006). Branch lengths in the species tree were then estimated using the ML with local clock and the above specified parameters of the Tamura-Nei model of substitutions with gamma distribution (Tamura and Nei, 1993) implemented in PAML v. 4.1 (Yang, 2007). In a case of the Bayesian phylogeny reconstruction, the maximum clade credibility tree was inferred using TreeAnnotator v.1.5.3, which is included in the BEAST software package.

Reconstruction of the evolutionary time scale

The inferred 16S-23S rRNA tree was tested for the presence of global and local molecular clock using the respective algorithms implemented in HYPHY (Kosakovsky-Pond et al., 2005). Based on the results of the test, the model with local clock was used for the further analysis.

Two internal calibration points were used for the time estimates. One point (CP1) is based on the fossil record about the origin of cyanobacteria and is constrained either by ~3500 MYA (Schopf and Packer, 1987; Walsh, 1992; Kazmierczak and Altermann, 2002; Altermann and Kazmierczak, 2003; Schopf et al., 2007), or by ~2600 MYA (Summons et al., 1999). Another point (CP2), which refers to the appearance of heterocystic cyanobacteria, was calibrated using both molecular and geological data and was estimated between 2450 and 2100 MYA (Tomitani et al., 2006). We used an average value of 2200 MYA for the analysis. These computations were conducted using PAML v. 4.1 (Yang, 2007). We also used the Bayesian relaxed clock phylogeny estimation (Drummond and Rambaut, 2007) as described above. Uncertainty in the estimates was indicated by 95% highest posterior density (95% HPD) intervals.

Modeling of the 3D structure of the GAF domain in the *bona fide* CikA protein

The GAF domain and the adjacent N-terminal region both are critical for the circadian function of CikA, as they control phosphorylation of the kinase domains (Mutsuda et al., 2003). Therefore, we modeled a 3D structure of the GAF domain and mapped on it the conserved motifs identified in this study. The 3D structure was modeled using a majority consensus sequence (in total 180 amino acids) from the alignment of the region between 20 *bona fide* CikA proteins. The initial model was constructed using (PS)² Protein Structure Prediction Server (<http://ps2.life.nctu.edu.tw/>) with following options selected: both PSI-BLAST (Altschul et al., 1997) and IMPALA (Schaffer et al., 1999) for template search, and RAMP (<http://software.compbio.washington.edu/ramp/ramp.html>) for the model building. The obtained initial model was then optimized using the MolProbity server at <http://molprobity.biochem.duke.edu/index.php> (Lovell et al., 2003). The quality of the

model's versions was assessed with PROCHECK (Laskowski et al., 1993), ERRAT (Colovos and Yeates, 1993), and VERIFY_3D (Luthy et al., 1992), as implemented in SAVES (http://nihserver.mbi.ucla.edu/SAVES_3/), and ProQ (Wallner and Elofsson, 2003) (<http://www.sbc.su.se/~bjornw/ProQ/ProQ.cgi>) The model built upon a consensus of the following three templates (PDB ID codes): 2K2N (Cornilescu et al., 2008), 2OOL (Yang et al., 2007) and 2O9C (Wagner et al., 2007) yielded the best scores.

Identification of amino acid residues of potential functional importance in the CikA proteins

The level of conservation is usually correlated to the functional importance of a particular amino acid site or a sequence motif (Kimura, 1983; Graur and Li, 2000). If particular sites in one protein subfamily are more conserved (or fixed) as compared to other subfamilies, they are assumed to be more functionally important for that subfamily. We applied ConSeq (Berezin et al., 2004) available at <http://conseq.tau.ac.il/index.html>) to identify which of the conserved residues are of potential functional significance.

Results

Architecture, occurrence, and phylogeny of the *cikA* genes and their homologs

Comparison of the *S. elongatus* CikA against the NCBI Conserved Domain Database (Marchler-Bauer et al., 2003) suggested that the protein consists of four tandemly arrayed functional domains: a GAF domain, a histidine kinase phosphoacceptor domain (HisKA), a histidine kinase-like ATPase domain (HATPase_c), and a signal receiver domain (REC) (Fig. 1). HisKA and HATPase_c are usually considered to be components of a single histidine-protein kinase (HPK)

unit and are referred to as dimerization and catalytic (ATP/ADP-binding phosphotransfer) domains, respectively (Stock, 1999). A BLAST search of available completed bacterial genomes revealed several hundred *cikA* homologs that may be classified as two-component histidine kinases. However, the majority of the matches were limited to either two (HisKA and HATPase_c) or three (HisKA, HATPase_c, and REC) domains. A relatively small number of the homologs possessed the GAF domain. Both GAF and its adjacent N-terminal region are critical for autophosphorylation of the HisKA domain and their deletion negatively affects CikA expression (Mutsuda et al., 2003). Therefore, only genes indicating homology to GAF as well as the HisKA and HATPase_c domains were selected for the further analyses. A number of the homologous genes containing other domains in addition to the originally described *cikA* four-domain architecture (GAF – HisKA – HATPase_c – REC) were found in cyanobacteria. The other domains included, but were not limited to, sensor domains (e.g., PAS, GAF, and the others), CBS, REC, HPT and the others. For example, the genes from the closely related thermophilic Yellowstone isolates *Synechococcus* sp. JA-2-3B'a(2-13) (JA-2 YP_476763) and *Synechococcus* sp. JA-3-3Ab (JA-3 YP_476201) (Table S1, Supplementary material online) have the following domain architecture: GAF – CHASE – PAS – PAS – GAF – HisKA – HATPase_c – REC – REC – HPT (Fig. S1, Supplementary material online).

Both ML and Bayesian phylogenetic analyses of the CikA-like proteins yielded an identical tree topology featuring five major distinct clades with high statistical support. They are designated hereafter as A1-A5 (Fig. 2). Clade A1 includes the originally described CikA from *S. elongatus* PCC 7942 (Schmitz et al., 2000) and its closest homologs, which are thus presumed to have a circadian function. These proteins usually have a fairly stable four-domain architecture GAF – HisKA –

HATPase_c – REC (as in CikA of *S. elongatus* PCC 7942), except for those of filamentous heterocystic Nostocales (genera *Anabaena* and *Nostoc*) lacking the REC domain (Fig. S1, Supplementary material online). All other CikA-like proteins are from cyanobacteria possessing the original *kaiABC* system, except for *Gloeobacter violaceus* PCC 7421, which does not have any *kai* genes (Nakamura et al., 2003). However, these homologs are more variable in their architecture, featuring various additional domains upstream in the N-terminal region or downstream in the C-terminal region, as described above (Fig. S1, Supplementary material online). The different domain architectures of the *cikA* homologs are likely indicative of their different functional assignments. Importantly, only the proteins from clade A1 (presumably the *bona fide* CikA proteins) occur in the genomes of all studied cyanobacterial species with the original *kaiABC* system. The proteins from clades A2-A5 were found only in a subset of the species presented here. These phylogenetic patterns are similar to those of other known circadian genes and their non-circadian homologs (Dvornyk, 2006b).

A notable feature of the *cikA* homologs is their high diversification in some species. The number of the homologs can vary considerably, from a single gene copy in *G. violaceus* PCC 7421, *Trichodesmium erythraeum* IMS101, and the others, up to the eight copies present in *Acaryochloris marina* MBIC11017 (Table S1, Supplementary material online).

Divergence of the *cikA* homologs and *cikA*-like genes

The *cikA* homologs from cyanobacteria exhibit domain-specific patterns of nucleotide variation. HisKA and HTPase_c are usually the most conserved domains (Table 1). The genes from A1, which presumably have a circadian function, appear to be the second least polymorphic after those of A2 (Table 1). Despite this, the REC domain of the genes from clade A1 displays the lowest level of conservation.

The CikA proteins of clade A1 have several regions that are much more conserved than in the other clades. These regions may be of particular importance for circadian functionality (Fig. 1 and Fig. S2, Supplementary material online). Specifically, one of these conserved motifs corresponds to the part of the N-terminal region immediately preceding the GAF domain (Fig. 3). The N-terminal region upstream of the GAF domain was previously shown to enhance phosphorylation of the HisKA domain (Mutsuda et al., 2003), however, no specific fragment of this region was identified as a major contributor to this function. The analysis of variation suggests that this fragment (motif 1) corresponds to amino acid residues 168-183 (numbering refers to positions in the CikA protein of *S. elongatus* PCC 7942). Notably, the whole N-terminal region preceding the GAF domain in CikA of *Lyngbya* sp. PCC 8106 and *Arthrospira maxima* CS-328 consists only of this motif (Fig. 3). This is further evidence for its functional significance. Another conserved region, motif 2, is located near motif 1 and includes residues 199-210 (Fig. 3).

The C-terminal region of the GAF domain exhibits a much lower level of variation than its other regions. Within its C-terminal section, a highly conserved segment of 10 residues (pos. 309-318, motif 3) was identified in the CikA proteins (Fig. 3). A search against the functional motif databases returned no apparent homologs of this motif, and its function remains undetermined.

In the *cikA* genes of clade A1, the highly conserved segments in the HisKA and HATPase_c domains are located near previously identified functionally important motifs (Schmitz et al., 2000; Mutsuda et al., 2003). For example, motif 4 (pos. 578-587) is immediately adjacent to the G-X-G motif (pos. 574-576) in the G box (Fig. S2, Supplementary material online). The G-X-G motifs are critical for ATP binding and are located in the loops that shape the top and bottom of the binding pocket (Obermann

et al., 1998). The segment between these motifs is more conserved in clade A1 than in the others, suggesting that the existing tertiary structure of the pocket is extremely important for the circadian function of the HATPase_c domain and CikA. Notably, the patterns of variation within the pocket highlight its different tertiary structure among the different clades.

The REC domain does not occur in some members of clade A1: it is missing in heterocystic filamentous Nostocaceae. Unlike the other three domains, REC is most polymorphic in clade A1, specifically at the functionally important fixed positions 678Asp, 708Thr, and 727Lys in clades A2 and A3. The conserved Asp residue functions as phosphoryl acceptor, while the conserved Lys is essential for the $\beta 5\alpha 5$ loop and formation of the dimer upon phosphorylation (Solá et al., 1999). Likewise, the intermolecular recognition site immediately following the phosphorylated residue (Müller-Dieckmann et al., 1999) is conserved in the other clades, while being polymorphic in A1.

The model of the CikA GAF domain

The quality assessment parameters of the constructed 3D model are presented in Table 2. They show that the predicted model is of good quality. The model basically follows the experimentally determined structure of the bacterial PAS-related domains (Wagner et al., 2007; Yang et al., 2007; Cornilescu et al., 2008): it consists of five antiparallel β -sheets located between two groups of α -helices (see Fig. S3, Supplementary material online). The β -sheets shape the bottom of a bilin binding pocket that is characteristic of GAF domains. The conserved motifs 1, 2, and 3 correspond to helix $\alpha 1$ and sheets $\beta 1$ and $\beta 5$, respectively.

Evolutionary constraints associated with functional specialization of the CikA proteins

The results of the Conseq analysis (Fig. S3, Supplementary material online) suggest that the conserved motifs identified in the CikA proteins may be functionally and/or structurally important. Motif 1 seems to be primarily functional: most of its residues are exposed and seven of them are determined as functionally important (Fig. S4). Motif 3 is mainly of structural significance, as all of its residues but one are buried. Motifs 2 and 4 are important both functionally and structurally, as they possess buried and exposed amino acid residues in about equal proportions. Furthermore, motif 4 is located near the ATP-binding pocket (Fig. 1, see above), which suggests the histidine in this position is of critical importance to CikA phosphorylation.

Evolutionary time estimates

We based our reconstruction of the time scale for the key events in evolution of CikA and the circadian system on the following facts (see Figs. 2 and 4 for reference): (1) *bona fide* CikA is present only in group S1; (2) the REC domain is missing in filamentous heterocystic Nostocaceae; (3) KaiA is missing in clade S3 (Dvornyk et al., 2003). These data constrain the time of the CikA origin to the period between nodes 1 and 2; the CikA loss in S2 and S3 – between nodes 3 and 5; the REC domain loss – between CP2 and node 4; the KaiA loss – between nodes 6 and 7. Table 3 presents the resulting time estimates for the respective nodes based on the results of the ML and Bayesian analyses. However, it should be taken into account that they are likely biased toward the higher values, because CP1 was placed in the node, which is apparently not basic for cyanobacteria, as other yet unknown cyanobacterial species evolutionarily older than *Gloeobacter* may potentially exist.

Discussion

CikA was identified as an important input component in the *kaiABC* system of *S. elongatus* PCC 7942 (Schmitz et al., 2000) and was hypothesized to transfer a signal to the central oscillator (presumably KaiA) through the yet unknown associated response regulator (Mutsuda et al., 2003; Zhang et al., 2006). However, as the results of the present study show, neither the *cikA* gene nor its apparent GAF-containing homologs occur in the cyanobacteria of clades S2 and S3 (Fig. 4). The species of clade S2 contain the *kaiA* gene, while those of clade S3 lack it. This leads to two conclusions. First, *cikA* was lost before *kaiA*. Second, if the species from either S2 or S3 possess circadian rhythmicity, they should utilize a different molecular mechanism for signal input. It was hypothesized earlier that information about the light signal may be delivered to the central oscillator not through a photoreceptor, but indirectly through sensing the redox state of the cell (Ivleva et al., 2006) by using another input element, such as LdpA (Ivleva et al., 2005). In contrast to CikA, LdpA occurs in all known cyanobacteria (Dvornyk, 2005; Dvornyk, unpublished]. This suggests that the indirect transfer of the light signal might have become a primary mechanism of the circadian input after the loss of *cikA*. The recent results of biochemical studies provide support for this hypothesis (Holtzendorff et al., 2008; Axmann et al., 2009). On the other hand, such an indirect signal transfer mechanism might have existed before the origin of *cikA*.

The circadian system of the species from clade S2 (Fig. 4) is of particular interest, because it shares structural features of the two systems: it has *kaiA* that makes it more similar to the *kaiABC* system, but lacks *cikA* as the *kaiBC* system does. In addition, other features of this circadian system (e.g., the evolutionary history of the *cpmA* gene regulating the circadian output) position it closer to the *kaiBC* type

(Dvornyk et al., 2003; Dvornyk et al., 2004; Dvornyk and Knudsen, 2005; Dvornyk, 2006b). If the *kaiABC*^A system of clade S2 is as functionally transitional between the *kaiABC* and *kaiBC* as it is structurally intermediate, then studying it may provide important information about the functional evolution of the original *kaiABC* system into the *kaiBC* system.

In the original study of CikA in *S. elongatus* PCC 7942, this protein was classified as a non-typical bacteriophytochrome, due to the missing conserved Cys residue at position 285 (Schmitz et al., 2000) that normally serves as a bilin ligand for phytochromes (Li and Lagarias, 1992). This suggests that this particular CikA may interact with another protein (possibly possessing the GAF domain) to replace the bilin acceptor (Mutsuda et al., 2003). However, given that most of the CikA proteins in clade A1 do possess this critical cysteine residue, the proposed mechanism may represent only one of the possible variants. Recent findings by Narikawa et al. (Narikawa et al., 2008) support this view, and show that GAF domains retaining the conserved C285 residue may function as a violet light sensor. This, in turn, explains the functional assignment of motifs 2 and 3 (Fig. 3), which ensure proper 3D configuration of the bilin-binding pocket (Fig. S3, Supplementary material online).

A recent study of CikA in *S. elongatus* PCC 7942 proposed positive regulation of CikA phosphorylation by the GAF domain and negative regulation by the REC domain (Mutsuda et al., 2003; Zhang et al., 2006). Autophosphorylation is essential for the circadian function of CikA (Mutsuda et al., 2003). This process involves all three principal domains of the protein (GAF as a positive regulator, HisKA as a phosphoacceptor, and HATPase_c as an ATP binder) and therefore assumes a close inter-domain interaction. For such an interaction, a corresponding tertiary structure of the CikA protein is critical for autophosphorylation. On the other hand, since CikA

belongs to the superfamily of sensor kinases of bacterial two-component signal transduction systems, it is expected to supply a phosphoryl group to a specific, yet unidentified, response regulator(s) (Schmitz et al., 2000; Mutsuda et al., 2003). This putative function of CikA assumes a close correspondence of its structure to that of the response regulator. The highly conserved motifs identified in the present study (Fig. 3) are probably critical for maintaining the tertiary structure of CikA, ensuring said physical interactions both between the domains and between each domain and the response regulator.

Recently, the pseudo-receiver REC domain was confirmed as necessary for the CikA circadian function in *S. elongatus* PCC 7942 by entraining the clock through sensing the redox state of cellular quinones (Ivleva et al., 2006) and repressing the kinase activity of the protein (Mutsuda et al., 2003; Zhang et al., 2006). However, REC is most polymorphic in clade A1 as compared to the other clades (Table 1) and is even missing in some cyanobacteria (e.g., Nostocales). According to the principle of functional constraint in molecular evolution, functional importance of a protein or domain directly correlates with its level of conservation (Kimura, 1983; Graur and Li, 2000). Thus the results of our study raise several questions about the functional significance of REC. Specifically, how is the protein autokinase activity of the REC-deficient CikA controlled? How is quinone sensing conducted? One possibility is that these REC-associated functions in some cyanobacterial species are performed by an unidentified response regulator protein.

Interestingly, two components of the cyanobacterial circadian system, CikA and SasA, which respectively control the input and output pathways in the *kaiABC* system, are autophosphorylated *in vitro* and have a similar domain structure (Sensor Domain – HisKA – HATPase_c) to the one that is common in two-component sensory

transduction histidine kinases (Dutta et al., 1999; Iwasaki et al., 2000; Mutsuda et al., 2003). The similar domain organization of CikA and SasA suggests that they both may have originated through the common evolutionary mechanism: the fusion of the sensor domain with a double-domain two-component sensory transduction histidine kinase. Two-component sensory transduction histidine kinases are a large superfamily widely distributed in prokaryotes (Nagaya et al., 1993). Proteins of this superfamily display a diversity of domain organizations (Dutta et al., 1999) that are the apparent result of multiple gene fusion events. Recent findings suggest gene fusions have played a major role in the evolution of protein-domain architectures (Yanai et al., 2002). These fusions were probably quite common in evolution of the CikA-like proteins. While the core domains, GAF, HiSKA and HTPase_c, exhibit relatively high similarity between the different members of clades A2-A5, the domain organization of the proteins varies greatly (Figs. 2 and S1, Supplementary material online).

The aggregated data of the current study and previous functional studies of *cikA* (Schmitz et al., 2000; Mutsuda et al., 2003; Zhang et al., 2006; Ivleva et al., 2006) provide evidence that various evolutionary mechanisms have resulted in functional specialization of *cikA* as a circadian gene. After the duplication of the ancestral two-component histidine kinase, *cikA* experienced neofunctionalization through accretion of specific domains (GAF and REC) while maintaining conservation at functionally important sites and domain architecture. While the CikA-like proteins from clades A2-A5 experienced significant diversification of the domain organization, the original CikA protein maintained its high level of conservation (Fig. S1, Supplementary material online). The acquired circadian function was then maintained by strong purifying selection in the regions conferring this function. This is a common pattern

for circadian genes, which are usually more conserved than their non-circadian paralogs (Dvornyk et al., 2004; Dvornyk, 2006b).

The results of the present study make it possible to roughly estimate the probable time of origin and the main events in the evolution of the *cikA* gene. The following time estimates are approximate, especially with regard to the very early stages of this evolution, and will be updated and corrected as additional genomic data are accumulating. According to the ML estimates, depending on the accepted date of the appearance of cyanobacteria, the lower time limit of *cikA* origin is placed about 2900 MYA, while the upper limit is placed about 2200 MYA (Table 3). The gene was lost in groups S2 and S3 in the period between 1050 and 600 MYA. The broad range of the estimate is due to the absence of data about the circadian system in cyanobacterial species that are phylogenetically positioned between *S. elongatus* PCC 7942 and *Synechococcus* sp. RCC307. The loss of the REC domain in filamentous Nostocales seems to have occurred before the loss of *cikA*, falling within the period of approximately 2200-900 MYA (Fig. 4, Table 3). The Bayesian analysis yielded similar to the above estimates for all nodes, except for nodes 3 and 5 (Table 3). The latter pull the date of the *cikA* loss back by almost two-fold, between 1900 and 1000 MYA. This large difference between the ML and Bayesian estimates is likely not due to the phylogenetic uncertainty, because both methods produced the same tree topology (Fig. 4). However, the most recent studies suggest that the ML analysis tends to give more accurate estimates of branch length and, respectively, divergence times (Schwartz and Mueller, 2010).

The *kaiABC* system was initially thought to have evolved from the *kaiBC* system through the addition of *kaiA*, which presumably originated about 1000 MYA (Dvornyk et al., 2003). The recent analysis of newly available, more extensive

genomic data suggests a different scenario: the current *kaiBC* system evolved stepwise during 1050-400 MYA from the *kaiABC* through the loss of *kaiA* and other components, including *cikA* (Dvornyk, 2006a; Dvornyk, 2009). Apparently, these losses were associated with the origin and rapid radiation of cyanobacteria in clades S2 and S3 (Fig. 4). Interestingly, the time limits of this radiation (between nodes 5 and 6, i.e., about 600-500 MYA, Fig.4 and Table 3) correspond to the period around the well-known Cambrian explosion. Furthermore, according to the results of the current study, *kaiA* was lost between 500 and 400 MYA (Table 3), which corresponds to the hypothesized upper time limit of the last of the three periods proposed to describe the role of UV radiation in the evolution of cyanobacteria (Garcia-Pichel, 1998). This period is thought to last between 1000 and 400 MYA and was associated with the increase of atmospheric oxygen content and the formation of the earth's ozone shield (Canfield and Teske, 1996; Garcia-Pichel, 1998). Recently, it was suggested that the loss of *kaiA* and the associated decrease in circadian oscillator robustness occurred due to the adaptation of *Prochlorococcus* to a temperature-stable ecological niche that does not require a robust oscillator (Holtzendorff et al., 2008; Axmann et al., 2009). This scenario may not fully explain the observed phylogenetic patterns, however, since *S. elongatus* PCC7942 possesses *kaiA* yet often occurs with *Prochlorococcus* in the same ecological niches (Partensky et al., 1999). Further studies may help to determine the factors that triggered the large-scale reduction of the *Prochlorococcus* genome (Dufresne et al., 2003) including the loss of several circadian genes.

An intriguing finding of our study is that none of the species from clades S2 and S3 have any GAF-containing genes. Obviously, these genes were present in the genome of the common ancestor of *S. elongatus* PCC 7942 and *Synechococcus* sp. RCC307 at the time point corresponding to node 3 (Fig. 4). However, for reasons yet

unknown, all GAF-containing genes were lost in the lineage leading to *Synechococcus* sp. RCC307. Of course, this loss might occur in any time point between nodes 3 and 5, but a more accurate estimate will only be possible when more genomic data in this lineage become available. The loss of *cikA* in some cyanobacterial lineages suggests that the evolution of this gene and its homologs follows the birth-and-death scenario (Nei and Rooney, 2005).

The results of the 16S-23S rRNA phylogenetic analysis provide some insights into molecular systematics of cyanobacteria. Specifically, assigning the species from clades S2 and S4 to the same genus *Synechococcus* seems unjustified. In fact, they are phylogenetically quite distant from each other as well as from *S. elongatus* PCC 7942. The polyphyly of the genus *Synechococcus* has been comprehensively discussed elsewhere (Urbach et al., 1998; Honda et al., 1999; Robertson et al., 2001). In total, up to eight *Synechococcus* lineages are recognized on the basis of 16S rRNA and other genes analyses (Honda et al., 1999; Robertson et al., 2001). However, this topic warrants a separate, more comprehensive investigation involving a much larger number of available strains.

All previous evolutionary studies have shown that the elements of the circadian system usually have lower variation than their apparent non-circadian homologs; this variation is specific to each type of the system and is maintained by strong purifying selection (Dvornyk et al., 2004; Dvornyk, 2005). While the core genes, *kaiB* and *kaiC*, are common among the different circadian system types, the input and output signal pathways differ significantly, thereby conferring functional and selective constraints to each type. The *kaiABC* system originally described in *S. elongatus* PCC 7942 is apparently evolutionarily oldest among the three; however it is still unclear what happened about 1050-600 MYA and 400 MYA that resulted in loss of *cikA* and *kaiA*,

respectively. Further comparative evolutionary and functional studies of all types of the cyanobacterial circadian system are needed to reveal the specific molecular mechanisms that have been developed and utilized during this system's evolution towards functional specialization.

Acknowledgements

This work was supported by the University of Hong Kong start-up fund for VD.

Abbreviations

cikA, circadian input kinase A; MYA, million years ago; NCBI, National Center for Biotechnology Information

Figures

Fig. 1. Domain architecture of the CikA protein with the mapped motifs of putative functional and/or structural importance. Dashed boxes represent regions not present in CikA of all species.

Fig. 2. Unrooted maximum-likelihood tree of the CikA homologs in cyanobacteria. Bar, 0.1 substitutions per site. Maximum-likelihood probabilities of the node support <0.5 and bootstrap <50 are not shown. A1-A5 refer to the subfamilies of the CikA homologs. Clade A1 comprises *bona fide* CikA proteins. For the designations of the proteins see Supplementary data file 1.

Fig. 3. Alignment of the GAF domains of the cyanobacterial CikA proteins. Motifs 1, 2, and 3 are underlined. Block arrow indicates putative cysteine ligand. Black and grey-shaded backgrounds indicate different degree of conservation (black is the most conserved). The upper numbers indicate positions in the alignment of the full sequences; the numbers on the right indicate positions in the respective sequences. Visualized using Genedoc (Nicholas and Nicholas, Jr., 1997).

Fig. 4. Maximum-likelihood tree with local clock of the concatenated 16S rRNA and 23S rRNA genes from cyanobacteria. Bar, 1 substitution per site. ML and posterior probabilities of the node support are shown in numerator, bootstrap proportion values are shown in denominator. ML values <0.5 and bootstrap <50 are not shown. ML and posterior probabilities values equal 1.00 are shown without decimals. S1-S4 refer to the groups of the species with different composition of the circadian system. Species with the *bona fide* CikA are boxed, those belonging to Nostocales and lacking the REC domain are shown in bold. Black boxes indicate the calibration points. Black circles and numbers mark nodes that correspond to the key events in evolution of CikA and the circadian system: the CikA origin – between nodes 1 and 2;

the CikA loss – between nodes 3 and 5; the REC domain loss – between CP2 and node 4; the KaiA loss – between nodes 6 and 7. See text for the details.

Tables

Table 1. Patterns of nonsynonymous nucleotide substitutions (d_N) in the different regions of the *cikA* genes and *cikA*-like homologs.

Clade	Domain					Average over gene ²
	GAF	HisKA	HTPase_c	REC ¹	Rest of gene	
A1	0.478 ± 0.039	0.272 ± 0.036	0.268 ± 0.032	0.495 ± 0.050	0.486 ± 0.039	0.417 ± 0.017
A2	0.229 ± 0.061	0.137 ± 0.063	0.274 ± 0.042	0.250 ± 0.040	0.614 ± 0.111	0.343 ± 0.051
A3	0.582 ± 0.115	0.453 ± 0.145	0.550 ± 0.105	–	0.879 ± 0.147	0.578 ± 0.088
A4	0.491 ± 0.039	0.260 ± 0.038	0.444 ± 0.046	0.446 ± 0.037	0.643 ± 0.055	0.484 ± 0.018
A5	0.615 ± 0.044	0.310 ± 0.047	0.451 ± 0.042	0.460 ± 0.068	0.685 ± 0.042	0.554 ± 0.022
Average over domain	0.767 ± 0.038	0.510 ± 0.043	0.566 ± 0.039	0.799 ± 0.055	0.920 ± 0.043	0.723 ± 0.025

¹ the number of the analyzed sequences is smaller due to the exclusion those lacking the REC domain;

² the REC domain is excluded as most genes from clade A2 lack it.

Table 2. Quality assessment parameters for the constructed 3D structure of the GAF domain.

Parameter	Value
Ramachandran plot	91.0% core, 7.2% allowed, 1.8% generously allowed, 0.0% disallowed
All Ramachandrans	8 labelled residues (out of 178)
Chi1-chi2 plots	1 labelled residues (out of 124)
Main-chain	6 better, 0 inside, 0 worse
Side-chain	5 better, 0 inside, 0 worse
Residue properties	Max.deviation: 2.3, bad contacts: 14, bond length/angle: 7.5, Morris et al class: 1 1 2
G-factors	Dihedrals: 0.06, covalent: -0.22, overall: -0.04
M/c bond lengths	98.7% within limits
M/c bond angles	92.8% within limits
Planar groups	100.0% within limits
Verify 3D score >0.2	65.19% of the residues
Errat overall quality factor	71.512
Predicted LGscore	4.477
Predicted MaxSub	0.267

Table 3. Maximum-likelihood and Bayesian time estimates, MYA, for the nodes (Fig. 4) corresponding to the major events in evolution of CikA and the circadian system. Based on the 16S-23S rRNA tree.

Node	Maximum-likelihood		Bayesian
	2600*	3500	3259 (2749-3793)**
1	2598 ± 15	2875 ± 219	3026 (2534-3515)
2	2223 ± 145	2217 ± 195	2583 (2186-2973)
3	1031 ± 105	1079 ± 111	1943 (1372-2524)
4	910 ± 63	952 ± 66	912 (400-1523)
5	588 ± 38	614 ± 40	1035 (576-1531)
6	499 ± 45	521 ± 48	670 (403-971)
7	407 ± 26	426 ± 28	565 (336-836)

* Time for calibration point 1, MYA,

** Posterior mean (95% HPD).

Supplementary material online

Supplementary data file 1 – List of sequences used in the study.

File format: pdf

Description: Tables S1, S2 and S3 provide a list of *cikA*, 16S rRNA and 23S rRNA sequences used in the study.

Supplementary Figure S1.

File format: pdf

Description: Phylogeny of the domain architectures of the *bona fide* CikA proteins and CikA-like proteins. Shown according to the clades in Fig. 1. Clades with the same domain architectures are collapsed.

Supplementary Figure S2.

File format: pdf

Description: Full multiple alignment of the *bona fide* CikA proteins. Black and grey-shaded backgrounds indicate different degree of conservation (black is the most conserved). The upper numbers indicate positions in the alignment of the full sequences; the numbers on the right indicate positions in the respective sequences. Visualized using Genedoc (Nicholas and Nicholas, Jr., 1997).

Supplementary Figure S3.

File format: pdf

Description: A theoretical 3D model of the GAF domain of the CikA protein. α -helices are shown in red, β -sheets are shown in yellow. A putative bilin-binding

residue, C285, which is conserved in most CikA proteins, is shown. Visualized with Sirius 1.2 (San Diego Supercomputer Center).

Supplementary Figure S4.

File format: pdf

Description: Predicted amino acid sites of functional and structural significance in the CikA protein as determined by ConSeq (Berezin et al., 2004). The predicted residues are mapped on the CikA protein of *S. elongatus* PCC 7942.

References

- Altermann W, Kazmierczak J (2003) Archean microfossils: a reappraisal of early life on Earth. *Res Microbiol* 154:611-617
- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* 215:403-410
- Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25:3389-3402
- Anisimova M, Gascuel O (2006) Approximate likelihood-ratio test for branches: A fast, accurate, and powerful alternative. *Syst Biol* 55:539-552
- Axmann IM, Duhring U, Seeliger L, Arnold A, Vanselow JT, Kramer A, Wilde A (2009) Biochemical evidence for a timing mechanism in *Prochlorococcus*. *J Bacteriol* 191:5342-5347
- Berezin C, Glaser F, Rosenberg J, Paz I, Pupko T, Fariselli P, Casadio R, Ben-Tal N (2004) ConSeq: the identification of functionally and structurally important residues in protein sequences. *Bioinformatics* 20:1322-1324
- Canfield DE, Teske A (1996) Late Proterozoic rise in atmospheric oxygen concentration inferred from phylogenetic and sulphur-isotope studies. *Nature* 382:127-132
- Colovos C, Yeates TO (1993) Verification of protein structures: patterns of nonbonded atomic interactions. *Protein Sci* 2:1511-1519
- Cornilescu G, Uliasz AT, Cornilescu CC, Markley JL, Vierstra RD (2008) Solution structure of a cyanobacterial phytochrome GAF domain in the red-light-absorbing ground state. *J Mol Biol* 383:403-413

- Drummond AJ, Rambaut A (2007) BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol Biol* 7:214
- Dufresne A, Salanoubat M, Partensky F, Artiguenave F, Axmann IM, Barbe V, Duprat S, Galperin MY, Koonin EV, Le Gall F, Makarova KS, Ostrowski M, Oztas S, Robert C, Rogozin IB, Scanlan DJ, Tandeau de Marsac N, Weissenbach J, Wincker P, Wolf YI, Hess WR (2003) Genome sequence of the cyanobacterium *Prochlorococcus marinus* SS120, a nearly minimal oxyphototrophic genome. *Proc Natl Acad Sci USA* 100:10020-10025
- Dutta R, Qin L, Inouye M (1999) Histidine kinases: diversity of domain organization. *Mol Microbiol* 34:633-640
- Dvornyk V (2005) Molecular evolution of *ldpA*, a gene mediating circadian input signal in cyanobacteria. *J Mol Evol* 60:105-112
- Dvornyk V (2006a) Evolution of the circadian clock mechanism in prokaryotes. *Isr J Ecol Evol* 52:343-357
- Dvornyk V (2006b) Subfamilies of *cpmA*, a gene involved in circadian output, have different evolutionary histories in cyanobacteria. *Microbiology* 152:75-84
- Dvornyk V (2009) The circadian clock gear in cyanobacteria: Assembled by evolution. In: Ditty, J. L., Mackey, S., and Johnson, C. H. (eds) *Bacterial Circadian Programs*. Springer-Verlag, Berlin-Heidelberg, pp 241-258
- Dvornyk V, Deng HW, Nevo E (2004) Structure and molecular phylogeny of *sasA* genes in cyanobacteria: insights into evolution of the prokaryotic circadian system. *Mol Biol Evol* 21:1468-1476
- Dvornyk V, Knudsen B (2005) Functional divergence of circadian clock proteins in prokaryotes. *Genetica* 124:247-254

- Dvornyk V, Vinogradova ON, Nevo E (2003) Origin and evolution of circadian clock genes in prokaryotes. *Proc Natl Acad Sci USA* 100:2495-2500
- Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 32:1792-1797
- Fankhauser C (2001) The phytochromes, a family of red/far-red absorbing photoreceptors. *J Biol Chem* 276:11453-11456
- Garcia-Pichel F (1998) Solar ultraviolet and the evolutionary history of cyanobacteria. *Orig Life Evol Biosph* 28:321-347
- Golden SS, Canales SR (2003) Cyanobacterial circadian clocks - timing is everything. *Nature Rev Microbiol* 1:191-199
- Graur D, Li WH (2000) *Fundamentals of Molecular Evolution*. Sinauer Associates, Sunderland, MA
- Guindon S, Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol* 52:696-704
- Holtzendorff J, Partensky F, Mella D, Lennon JF, Hess WR, Garczarek L (2008) Genome streamlining results in loss of robustness of the circadian clock in the marine cyanobacterium *Prochlorococcus marinus* PCC 9511. *J Biol Rhythms* 23:187-199
- Honda D, Yokota A, Sugiyama J (1999) Detection of seven major evolutionary lineages in cyanobacteria based on the 16S rRNA gene sequence analysis with new sequences of five marine *Synechococcus* strains. *J Mol Evol* 48:723-739
- Ishiura M, Kutsuna S, Aoki S, Iwasaki H, Andersson CR, Tanabe A, Golden SS, Johnson CH, Kondo T (1998) Expression of a gene cluster *kaiABC* as a circadian feedback process in cyanobacteria. *Science* 281:1519-1523

- Ivleva NB, Bramlett MR, Lindahl PA, Golden SS (2005) LdpA: a component of the circadian clock senses redox state of the cell. *EMBO J* 24:1202-1210
- Ivleva NB, Gao T, LiWang AC, Golden SS (2006) Quinone sensing by the circadian input kinase of the cyanobacterial circadian clock. *Proc Natl Acad Sci USA* 103:17468-17473
- Iwasaki H, Williams SB, Kitayama Y, Ishiura M, Golden SS, Kondo T (2000) A *kaiC*-interacting sensory histidine kinase, SasA, necessary to sustain robust circadian oscillation in cyanobacteria. *Cell* 101:223-233
- Kazmierczak J, Altermann W (2002) Neoarchean biomineralization by benthic cyanobacteria. *Science* 298:2351
- Kimura M (1983) *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge
- Kitayama Y, Iwasaki H, Nishiwaki T, Kondo T (2003) KaiB functions as an attenuator of KaiC phosphorylation in the cyanobacterial circadian clock system. *EMBO J* 22:2127-2134
- Kondo T, Ishiura M (1999) The circadian clocks of plants and cyanobacteria. *Trends Plant Sci* 4:171-176
- Kosakovsky-Pond SL, Frost SD, Muse SV (2005) HyPhy: hypothesis testing using phylogenies. *Bioinformatics* 21:676-679
- Laskowski RA, MacArthur MW, Moss DS, Thornton JM (1993) PROCHECK: a program to check the stereochemical quality of protein structures. *J Appl Cryst* 26:283-291
- Le SQ, Gascuel O (2008) An improved general amino acid replacement matrix. *Mol Biol Evol* 25:1307-1320

- Li L, Lagarias JC (1992) Phytochrome assembly. Defining chromophore structural requirements for covalent attachment and photoreversibility. *J Biol Chem* 267:19204-19210
- Li W-H (1993) Unbiased estimation of the rates of synonymous and nonsynonymous substitution. *J Mol Evol* 36:96-99
- Lovell SC, Davis IW, Arendall WB, III, de Bakker PI, Word JM, Prisant MG, Richardson JS, Richardson DC (2003) Structure validation by C α geometry: ϕ , ψ and C β deviation. *Proteins* 50:437-450
- Luthy R, Bowie JU, Eisenberg D (1992) Assessment of protein models with three-dimensional profiles. *Nature* 356:83-85
- Marchler-Bauer A, Anderson JB, DeWeese-Scott C, Fedorova ND, Geer LY, He S, Hurwitz DI, Jackson JD, Jacobs AR, Lanczycki CJ, Liebert CA, Liu C, Madej T, Marchler GH, Mazumder R, Nikolskaya AN, Panchenko AR, Rao BS, Shoemaker BA, Simonyan V, Song JS, Thiessen PA, Vasudevan S, Wang Y, Yamashita RA, Yin JJ, Bryant SH (2003) CDD: a curated Entrez database of conserved domain alignments. *Nucleic Acids Res* 31:383-387
- Müller-Dieckmann HJ, Grantz AA, Kim SH (1999) The structure of the signal receiver domain of the *Arabidopsis thaliana* ethylene receptor ETR1. *Structure* 7:1547-1556
- Mutsuda M, Michel KP, Zhang X, Montgomery BL, Golden SS (2003) Biochemical properties of CikA, an unusual phytochrome-like histidine protein kinase that resets the circadian clock in *Synechococcus elongatus* PCC 7942. *J Biol Chem* 278:19102-19110

- Nagaya M, Aiba H, Mizuno T (1993) Cloning of a sensory-kinase-encoding gene that belongs to the two-component regulatory family from the cyanobacterium *Synechococcus* sp. PCC7942. *Gene* 131:119-124
- Nakamura Y, Kaneko T, Sato S, Mimuro M, Miyashita H, Tsuchiya T, Sasamoto S, Watanabe A, Kawashima K, Kishida Y, Kiyokawa C, Kohara M, Matsumoto M, Matsuno A, Nakazaki N, Shimpo S, Takeuchi C, Yamada M, Tabata S (2003) Complete genome structure of *Gloeobacter violaceus* PCC 7421, a cyanobacterium that lacks thylakoids. *DNA Res* 10:137-145
- Narikawa R, Kohchi T, Ikeuchi M (2008) Characterization of the photoactive GAF domain of the CikA homolog (SyCikA, Slr1969) of the cyanobacterium *Synechocystis* sp. PCC 6803. *Photochem Photobiol Sci* 7:1253-1259
- Nei M, Rooney AP (2005) Concerted and birth-and-death evolution of multigene families. *Annu Rev Genet* 39:121-152
- Nicholas, K. B. and Nicholas, H. B., Jr. GeneDoc: a tool for editing and annotating multiple sequence alignments. 1997
- Obermann WM, Sondermann H, Russo AA, Pavletich NP, Hartl FU (1998) In vivo function of Hsp90 is dependent on ATP binding and ATP hydrolysis. *J Cell Biol* 143:901-910
- Pamilo P, Bianchi NO (1993) Evolution of the *Zfx* and *Zfy* genes: Rates and interdependence between the genes. *Mol Biol Evol* 10:271-281
- Partensky F, Hess WR, Vaulot D (1999) *Prochlorococcus*, a marine photosynthetic prokaryote of global significance. *Microbiol Mol Biol Rev* 63:106-127
- Posada D, Crandall KA (1998) MODELTEST: testing the model of DNA substitution. *Bioinformatics* 14:817-818

- Robertson BR, Tezuka N, Watanabe MM (2001) Phylogenetic analyses of *Synechococcus* strains (cyanobacteria) using sequences of 16S rDNA and part of the phycocyanin operon reveal multiple evolutionary lines and reflect phycobilin content. *Int J Syst Evol Microbiol* 51:861-871
- Schaffer AA, Wolf YI, Ponting CP, Koonin EV, Aravind L, Altschul SF (1999) IMPALA: matching a protein sequence against a collection of PSI-BLAST-constructed position-specific score matrices. *Bioinformatics* 15:1000-1011
- Schmitz O, Katayama M, Williams SB, Kondo T, Golden SS (2000) CikA, a bacteriophytochrome that resets the cyanobacterial circadian clock. *Science* 289:765-768
- Schopf JW, Kudryavtsev AB, Czaja AD, Tripathi AB (2007) Evidence of archaean life: Stromatolites and microfossils. *Precambrian Res* 158:141-155
- Schopf JW, Packer BM (1987) Early Archaean (3.3-billion to 3.5-billion-year-old) microfossils from Warrawoona Group, Australia. *Science* 237:70-73
- Schwartz RS, Mueller RL (2010) Branch length estimation and divergence dating: estimates of error in Bayesian and maximum likelihood frameworks. *BMC Evol Biol* 10:5
- Solá M, Gomis-Rüth FX, Serrano L, González A, Coll M (1999) Three-dimensional crystal structure of the transcription factor PhoB receiver domain. *J Mol Biol* 285:675-687
- Stock J (1999) Signal transduction: Gyrate protein kinases. *Curr Biol* 9:R364-R367
- Summons RE, Jahnke LL, Hope JM, Logan GA (1999) 2-Methylhopanoids as biomarkers for cyanobacterial oxygenic photosynthesis. *Nature* 400:554-557
- Tamura K, Dudley J, Nei M, Kumar S (2007) MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Mol Biol Evol* 24:1596-1599

- Tamura K, Nei M (1993) Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol Biol Evol* 10:512-526
- Tomitani A, Knoll AH, Cavanaugh CM, Ohno T (2006) The evolutionary diversification of cyanobacteria: molecular-phylogenetic and paleontological perspectives. *Proc Natl Acad Sci USA* 103:5442-5447
- Urbach E, Scanlan DJ, Distel DL, Waterbury JB, Chisholm SW (1998) Rapid diversification of marine picophytoplankton with dissimilar light-harvesting structures inferred from sequences of *Prochlorococcus* and *Synechococcus* (Cyanobacteria). *J Mol Evol* 46:188-201
- Wagner JR, Zhang J, Brunzelle JS, Vierstra RD, Forest KT (2007) High resolution structure of *Deinococcus* bacteriophytochrome yields new insights into phytochrome architecture and evolution. *J Biol Chem* 282:12298-12309
- Wallner B, Elofsson A (2003) Can correct protein models be identified? *Protein Sci* 12:1073-1086
- Walsh MM (1992) Microfossils and possible microfossils from the Early Archean Onverwacht Group, Barberton Mountain Land, South Africa. *Precambrian Res* 54:271-293
- Wernersson R, Pedersen AG (2003) RevTrans: Multiple alignment of coding DNA from aligned amino acid sequences. *Nucleic Acids Res* 31:3537-3539
- Whelan S, Goldman N (2001) A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol Biol Evol* 18:691-699

- Woelfle MA, Ouyang Y, Phanvijhitsiri K, Johnson CH (2004) The adaptive value of circadian clocks: an experimental assessment in cyanobacteria. *Curr Biol* 14:1481-1486
- Xu Y, Mori T, Johnson CH (2003) Cyanobacterial circadian clockwork: roles of KaiA, KaiB and the *kaiBC* promoter in regulating KaiC. *EMBO J* 22:2117-2126
- Yanai I, Wolf YI, Koonin EV (2002) Evolution of gene fusions: horizontal transfer versus independent events. *Genome Biol* 3:RESEARCH0024
- Yang X, Stojkovic EA, Kuk J, Moffat K (2007) Crystal structure of the chromophore binding domain of an unusual bacteriophytochrome, RpBphP3, reveals residues that modulate photoconversion. *Proc Natl Acad Sci USA* 104:12571-12576
- Yang Z (2007) PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* 24:1586-1591
- Zhang X, Dong G, Golden SS (2006) The pseudo-receiver domain of CikA regulates the cyanobacterial circadian input pathway. *Mol Microbiol* 60:658-668

Figure 1



Figure 2

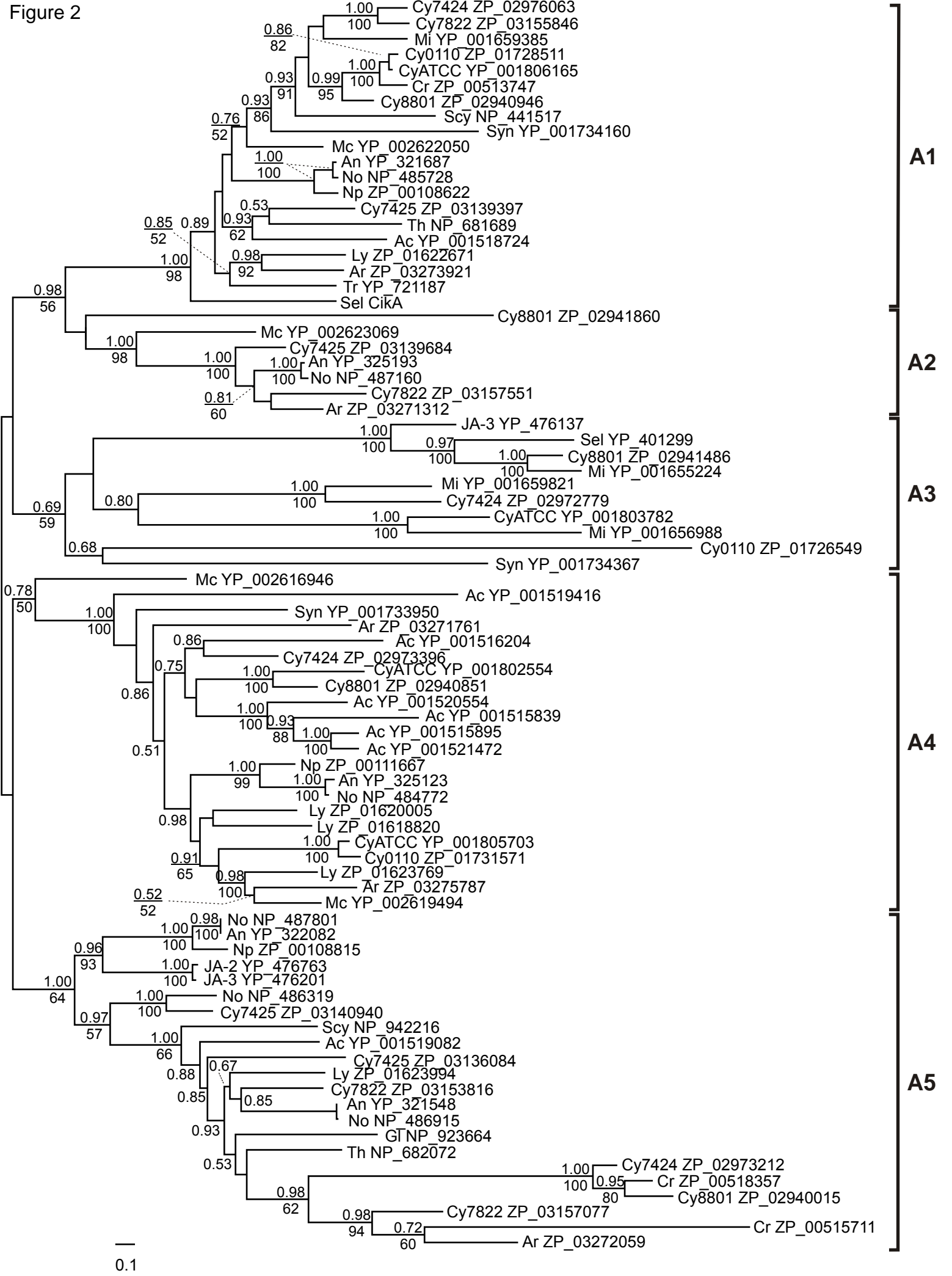


Figure 3

		*	360	*	380	*	400	*	420	*	440	*
Sel_Cika	:	-----	YPAVSVCQPLEQALHWQEE	----	QDRLSQVSAQIRLSLDLSHLTTITREIRQLLNADRAIIYQFKPQCCLDAGLDQRWPLYIP	-----	:	228				
Ac_YP_001518724	:	-----	AVPLA-----	PEVTLHPTAN	----	QERLSHQIISLSQGLPALLSMVVEIRHLQADRLIIY	-----	QLSPSTVNGAEASHHPVSA	-----	:	214	
An_YP_321687	:	-----	SEEVYAQCAVEDALKKQIS	----	QERLINQVTTQIRKSLDLPVIMATAIAQVREFLELDRLVIYKFEGRVNTQTRPPRLEDWQN	-----	:	283				
Ar_ZP_03273921	:	-----	MDDSLNKOLE	----	QERLINQVTTQIRQSLDLSVILSTAVQQLREFLQVDRLLVIYQLNSYQVQTGDHTKDLSPDVDEPDLNTEIPVPQADRHQD	-----	:	90				
Cr_ZP_00513747	:	-----	YEGRLTCEPVEALRQOVA	----	QERLINQLIGQIHQSLELSVILETAVREIRSFLEVDRLIIYKFKFKYLTSTSESKIQN	-----	:	229				
Cy0110_ZP_01728511	:	-----	YEGRLVCEPVEDALRQQA	----	QEKLLNQVVGQIRQSLSELSVILQTAVREIRSFLOVDRLLVIYEFKFKFKIISNSSSRNSRQ	-----	:	232				
Cy7424_ZP_02976063	:	-----	SPPTSLYHPIEALRQOVE	----	QERLINQVVTQIHQSLDLSQVLKTAVTEVNELGVDRLLIYQKFNSSVTDSTPHSLT	-----	:	224				
Cy7425_ZP_03139397	:	-----	TTVSLNSPLSPW	----	QQLLTOMTTOIWQNDLSLILQTTLEQVQRVLEVDRLIIYQLQ	-----	:	204				
Cy7822_ZP_03155846	:	-----	CHASVCQPIEALRQOVE	----	QERLINQVITQIRHSLELPVILETAVTEVNELQVDRLLIYQFSSHPSETETKTQTFPN	-----	:	229				
Cy8801_ZP_02940946	:	-----	YTTAMVCKPVEDALRQQA	----	QERLINQVIAQIRQSLNLSAILKTAVREVRSFLOVDRLLVIYEFQGQTPPNSEYSQFFT	-----	:	229				
CyATCC_YP_001806165	:	-----	YEGRLVCEPVEDALRQOVA	----	QERLINQVVGQIRQSLSELSVILETAVREIRSFLOVDRLLVIYEFREKTIISNNQSENAKQ	-----	:	233				
Ly_ZP_01622671	:	-----	MDKALDQOLE	----	QERLINQVTSQIRQSLSELSLILSTHVEQLQQFLQVDRLLIYQDFQWSPISPRECSLKPRLDLEEQNRDTRLTPPLPKDSSRSKPS	-----	:	95				
Mc_YP_002622050	:	-----	IEALRQOVE	----	QERLHVQVTHTHIRQSLLELPVILSTAVDEVHFLNVDRLIIYQFDGDLVNLRSEGDKIEVDNSTDVTDEFENAQFLTSSPLSQ	-----	:	260				
Mi_YP_001659385	:	-----	SRIFSVQCPVEAALSQQIQ	----	QERLINQVITQMRQSLLEPVILETVREAREFLQVDRLLIYQFFPSTSE	-----	:	210				
No_NP_485728	:	-----	SEEVSYCAVEDALKKQIS	----	QERLINQVTTQIRKSLDLPVIMATAIAQVREFLELDRLVIYKFEASKVKTQEYQSSTDEBNGKGSTISVNNQSLEDYQQ	-----	:	283				
Np_ZP_00108622	:	-----	APAVYIQCPVEDALKKQIS	----	QERLINQVTTQIRKSLDLPVIMATAIAQVREFLELDRLVIYKFEASKVKTQEYQSSTDEBNGKGSTISVNNQSLEDYQQ	-----	:	295				
Scy_NP_441517	:	-----	PEPSLQDSIQAS	----	QVKLSQVIAQIRQSLDLSHLNNAVTAQKELFVDRLLVIYQFHYSQPSLTPLEENQIPAPRPRQ	-----	:	222				
Syn_YP_001734160	:	-----	DEVVVPNQPLNMLLSQRL	----	QEKIINNVTTHRIYQNDLMVITVRMALEQAQRLLRVDRLLVIYQLDLPTAKPEK	-----	:	227				
Th_NP_681689	:	-----	TAQCQSQCPVTALNLRQA	----	QERLHVQVTTQIRQSLLELPILKIAVDRIREFLDVDRLLVGFQAQTEGE	-----	:	201				
Tr_YP_721187	:	-----	SDSVMDQEYPLVCKPIETALNQLE	----	QERLINQVTAQIRQSLLELPILSLAVKEVRKFIQVDRLLIYEINNPILLTDNITTVNE	-----	:	322				

```

Sel_Cika      : ---SQSYTHYEDRRNEALPSVIDPLVQP-GLLI-TTEWQRFQOGETLILDSGFYKERLPEQYSFYERVQVRSVKXIPILVQGRIGWGLLVAHQCCQDHRWQPRRERDILQHHA : 336
Ac_YP_001518724 : ---QSGSVVHEALASLPSVLV--DQEQLWGFIASSELSSLYQGYTHAIHVDLQNDVSSEHAKVNSIOVRALVITPILVKSLEWGLLIAHQCHLHPROMQVQEQDILLKRA : 322
An_YP_321687    : ---YGCIVYEAREATDILPSVLDY-QEKTFSRHS-QWEKYRQOFTIIVDDIEQAALAEELCNLFNRESQVRAKIAAPIIFEDKLWGLLIAHQCHSPROWNDSDKNLLISTA : 391
Ar_ZP_03273921  : ---IHGSTIYEAKSSNATASLLDWEAADRCFAQVP-EYQYQYKSIITHCIPNKLKYQSSPEYYSIQQKQVTAQLVAPILVQKQLWGILIAHQCFSDROWKESEKQFISKIIT : 199
Cr_ZP_00513747  : ---LGGLVYIESRVSQSLPSLLNVVAEDVCFATIP-EYQHKKYRQGEIYAIEDVERQSSASLCSQFTEKYWLSKLIAPILVEGELWGLLIAHQCFKKROWLESEKAFLLGQIG : 338
Cy0110_ZP_01728511 : ---TGGLVYIESRVSQSLPSLLNVVAEDTCFTNIP-QYKYKQRQGIIVSIDDVEMRAASLCLSRFLEQYVWLSKLIAPILVNGKLWGLLIAHQCFKKROWLESEKAFLLGQIG : 341
Cy7424_ZP_02976063 : ---RKATIIYEARSQSLPSMLHLSPEEDCFVCP-FCYQKYRQGTVAVENAETESSSFCIAELIRQNHVSKLIIAPILVDEQLWGLLIAHQCFKIROWLESEKGFILGHIG : 333
Cy7425_ZP_03139397 : ---THHQITHEALAAHAPSLILG-KGDPITCLHST-ACYQRYLQGEIVQVSDVEATADTPCLLALIQSQVRAKLIIPILVQEQELWGLLIAHQDPIROWQPPQEQQLGQIG : 312
Cy7822_ZP_03155846 : ---GWCKIIEARASQLPSLLNMPEDDCFSYLP-QYQDKYRRGTVAVENVETQSSSFCITEFIRHNQVLSLLIAPILVDEDQLWGLLIAHQCFKKROWLETEKEFLGHIG : 338
Cy8801_ZP_02940946 : ---SWGCVYIESKVSNSLPSLLNVVAEDCCFSHIP-HYQEKYRQGAIVIEDVEEAASSFCINLFEKYVIRAKLIAPILVENKLWGLLIAHQCFNKRWFSDSEKNFLGQIG : 338
CyATCC_YP_001806165 : ---TSLGVYIESRVSQSLPSLLNLVAEDTCFSNIP-QYKYKQRQGIIVSIDDVEMRAASLCLSRFLEKYWLSKLIAPILVNGKLWGLLIAHQCFKKROWLESEKAFLLGQIG : 342
Ly_ZP_01622671  : EKGYGCVYIEARSSENILSVLHINETDSCFFDTPKQAQQTQNSQALIMQDVDAQVDSPLFLEFHQIQVRSKIVFSIVQDQLWGLLIAHQCSQLRSWTDSETKFIARIT : 208
Mc_YP_002622050 : ---DEGCVYIEARRSNEIRSVLNWREKDCALYVS-NCRSKYHQGFTGAIDIEVAISDSPCLQLLRQTOYKAKIVAPILVQNRLLWGLLIAHQCFAPRHWOSEKFLFKQITA : 369
Mi_YP_001659385 : ---VKGKITSESRATQINSVLNLTPEDDCFYSYIP-QYKEKYRQGLIIVAVDVIDANSSSFCISEFIRHQHVSQSKLIAPILVQEEELWGLLIAHQCFEQRONLPPQEKFLGQIG : 319
No_NP_485728    : ---YGCIVYEAREATDILPSVLDY-QEQTFSRHS-QWEDKYRQOFTIIVDDIEQAALAEELCNLFNRESQVRAKIAAPIIFEDKLWGLLIAHQCYNPROWHDSKNLLISTA : 391
Np_ZP_00108622  : ---HGHYIYVEARAIDAITCLVLY-TEKNCFMRTS-QCWEKYRQOFTIIVDDVEKTALEELCNLFNRESQVRAKIAAPIITENKLWGLLIAHQQNEPROWDSEKNLLIAHA : 403
Scy_NP_441517   : ---QYGEVYIEARRSPEIDTMLGIMTENDCFSQVF-SYQKYIKGAVAVSDIENHSSSYCHVGLIQRQYVRAKIVAPILVEGQLWGLLIAHQCHHPROWLSDSEKNFLGQIG : 331
Syn_YP_001734160 : ---FVNRVYIEVVSDDKVLHLHF-QDDBCLTNE-LKIDAYLEGGKHAIVNDITQPNLSLFCFKQQLQGMQWCAKIVPLVILVEQRLWGLLIAHQCHGPRFRKNETIFLSHVA : 335
Th_NP_681689    : ---LRGQIYIESCRNSEPSVLGWDCCDQWMSGLPSSSYQRLSQGEAIVVSDITQQFYGAVPCHQSFAAHWCIKSWLIVPILVQDKLWGLLIAHQDRPROWQDQVEFLTHLS : 311
Tr_YP_721187    : ---QKGSVYIESLVSNTKISLLGLSERKNCFILPNSSIFYEYSKIS-VHFIADIEVAEKKFPELELEHQHCIAKALIVPILVQDKLWGLLIAHQCSEREWQDDKEREFLLETA : 430

```


1

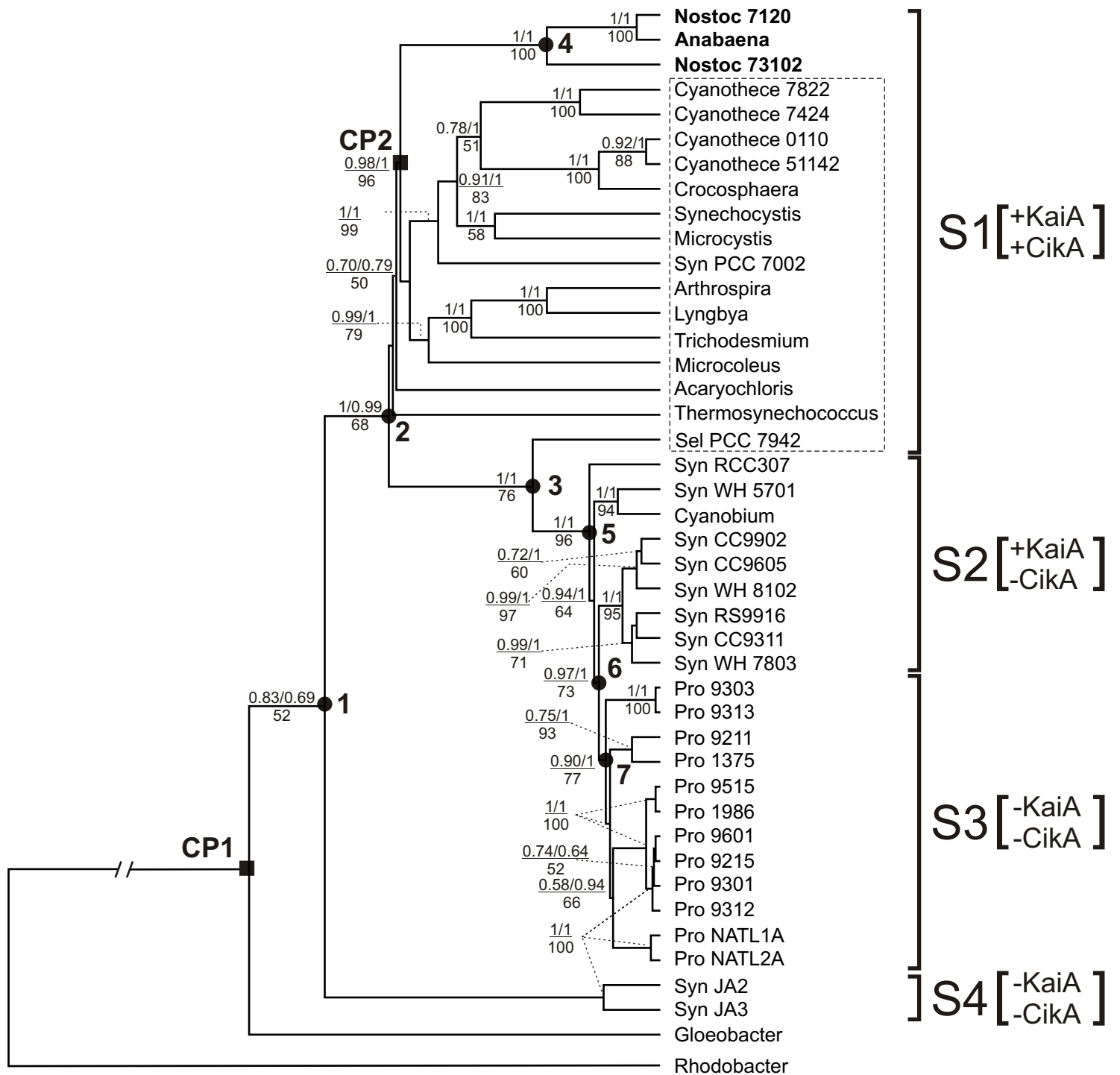


Table S1 – List of the *cikA* sequences used in the study.

Species and strain	Length, bp	GenBank nucleotide sequence accession number	Protein accession number	Designation of sequence in figure
<i>Acaryochloris marina</i> MBIC11017	2208	NC_009925	YP_001518724	Ac YP_001518724
	2790	NC_009925	YP_001515839	Ac YP_001515839
	4545	NC_009925	YP_001516204	Ac YP_001516204
	2667	NC_009925	YP_001519082	Ac YP_001519082
	2652	NC_009925	YP_001515895	Ac YP_001515895
	2715	NC_009925	YP_001520554	Ac YP_001520554
	2790	NC_009928	YP_001521472	Ac YP_001521472
	6123	NC_009925	YP_001519416	Ac YP_001519416
<i>Anabaena variabilis</i> ATCC 29413	2031	NC_007413	YP_321687	An YP_321687
	3825	NC_007413	YP_325193	An YP_325193
	4653	NC_007413	YP_322082	An YP_322082
	5463	NC_007413	YP_321548	An YP_321548
	5019	NC_007413	YP_325123	An YP_325123
<i>Arthrospira maxima</i> CS-328	1878	NZ_ABYK01000018	ZP_03273921	Ar ZP_03273921
	5310	NZ_ABYK01000001	ZP_03271312	Ar ZP_03271312
	3156	NZ_ABYK01000049	ZP_03275787	Ar ZP_03275787
	4176	NZ_ABYK01000005	ZP_03272059	Ar ZP_03272059
	4101	NZ_ABYK01000003	ZP_03271761	Ar ZP_03271761
<i>Crocospaera watsonii</i> WH 8501	2277	NZ_AADV02000001	ZP_00513747	Cr ZP_00513747

	2391	NZ_AADV02000007	ZP_00515711	Cr ZP_00515711
	1707	NZ_AADV02000134	ZP_00518357	Cr ZP_00518357
<i>Cyanothece</i> sp. ATCC 51142	2298	NC_010546	YP_001806165	CyATCC YP_001806165
	3255	NC_010546	YP_001802554	CyATCC YP_001802554
	2595	NC_010546	YP_001805703	CyATCC YP_001805703
	2235	NC_010546	YP_001803782	CyATCC YP_001803782
<i>Cyanothece</i> sp. CCY0110	2295	NZ_AAXW01000009	ZP_01728511	Cy0110 ZP_01728511
	2529	NZ_AAXW01000057	ZP_01731571	Cy0110 ZP_01731571
	2697	NZ_AAXW01000002	ZP_01726549	Cy0110 ZP_01726549
<i>Cyanothece</i> sp. PCC 7424	2274	NZ_ABOY01000027	ZP_02976063	Cy7424 ZP_02976063
	3858	NZ_ABOY01000007	ZP_02973396	Cy7424 ZP_02973396
	1734	NZ_ABOY01000006	ZP_02973212	Cy7424 ZP_02973212
	1839	NZ_ABOY01000004	ZP_02972779	Cy7424 ZP_02972779
<i>Cyanothece</i> sp. PCC 7425	2202	NZ_ABVJ01000010	ZP_03139397	Cy7425 ZP_03139397
	5934	NZ_ABVJ01000012	ZP_03139684	Cy7425 ZP_03139684
	3918	NZ_ABVJ01000022	ZP_03140940	Cy7425 ZP_03140940
	2664	NZ_ABVJ01000001	ZP_03136084	Cy7425 ZP_03136084
<i>Cyanothece</i> sp. PCC 7822	2289	NZ_ABVE01000004	ZP_03155846	Cy7822 ZP_03155846
	4461	NZ_ABVE01000008	ZP_03157551	Cy7822 ZP_03157551
	4281	NZ_ABVE01000001	ZP_03153816	Cy7822 ZP_03153816
	4842	NZ_ABVE01000006	ZP_03157077	Cy7822 ZP_03157077
<i>Cyanothece</i> sp. PCC 8801	2292	NZ_ABLR01000005	ZP_02940946	Cy8801 ZP_02940946
	2535	NZ_ABLR01000004	ZP_02940851	Cy8801 ZP_02940851

	1329	NZ_ABLR01000008	ZP_02941860	Cy8801 ZP_02941860
	1731	NZ_ABLR01000002	ZP_02940015	Cy8801 ZP_02940015
	2076	NZ_ABLR01000006	ZP_02941486	Cy8801 ZP_02941486
<i>Gloeobacter violaceus</i> PCC 7421	2571	NC_005125	NP_923664	Gl NP_923664
<i>Lyngbya</i> sp. PCC 8106	1926	NZ_AAVU01000031	ZP_01622671	Ly ZP_01622671
	3678	NZ_AAVU01000006	ZP_01620005	Ly ZP_01620005
	3084	NZ_AAVU01000048	ZP_01623769	Ly ZP_01623769
	4884	NZ_AAVU01000001	ZP_01618820	Ly ZP_01618820
	5184	NZ_AAVU01000054	ZP_01623994	Ly ZP_01623994
<i>Microcoleus chthonoplastes</i> PCC 7420	2370	NW_002435256.1	YP_002622050	Mc YP_002622050
	5568	NW_002435259.1	YP_002623069	Mc YP_002623069
	4425	NW_002435233.1	YP_002616946	Mc YP_002616946
	3069	NW_002435248.1	YP_002619494	Mc YP_002619494
<i>Microcystis aeruginosa</i> NIES-843	2196	NC_010296	YP_001659385	Mi YP_001659385
	2238	NC_010296	YP_001656988	Mi YP_001656988
	1992	NC_010296	YP_001655224	Mi YP_001655224
	2229	NC_010296	YP_001659821	Mi YP_001659821
<i>Nostoc punctiforme</i> PCC 73102	2052	NZ_AAAAY02000045	ZP_00108622	Np ZP_00108622
	5667	NZ_AAAAY02000005	ZP_00111667	Np ZP_00111667
	4680	NZ_AAAAY02000034	ZP_00108815	Np ZP_00108815
<i>Nostoc</i> sp. PCC 7120	2031	NC_003272	NP_485728	No NP_485728
	3861	NC_003272	NP_487160	No NP_487160
	3900	NC_003272	NP_486319	No NP_486319

	4938	NC_003272	NP_484772	No NP_484772
	4653	NC_003272	NP_487801	No NP_487801
	5454	NC_003272	NP_486915	No NP_486915
<i>Synechococcus elongatus</i> PCC 7942	2265	NC_007604	YP_399663	Sel CikA
	1914	NC_007604	YP_401299	Sel YP_401299
<i>Synechococcus</i> sp. JA-2-3B'a(2-13)	4581	NC_007776	YP_476763	JA-2 YP_476763
<i>Synechococcus</i> sp. JA-3-3Ab	4572	NC_007775	YP_476201	JA-3 YP_476201
	1965	NC_007775	YP_476137	JA-3 YP_476137
<i>Synechococcus</i> sp. PCC 7002	2283	NC_010475	YP_001734160	Syn YP_001734160
	2325	NC_010475	YP_001733950	Syn YP_001733950
	2826	NC_010475	YP_001734367	Syn YP_001734367
<i>Synechocystis</i> sp. PCC 6803	2253	NC_000911	NP_441517	Scy NP_441517
	3630	NC_005229	NP_942216	Scy NP_942216
<i>Thermosynechococcus elongatus</i> BP-1	2190	NC_004113	NP_681689	Th NP_681689
	4062	NC_004113	NP_682072	Th NP_682072
<i>Trichodesmium erythraeum</i> IMS101	2562	NC_008312	YP_721187	Tr YP_721187

* identical to the sequence of *S. elongatus* PCC 7942

Table S2 – List of 16S rRNA sequences used in the study.

Species and strain	Length, bp	GenBank nucleotide sequence accession number	GI/EMB List	Designation of sequence in the figure [*]
<i>Acaryochloris marina</i> MBIC11017	1501	NC_009925	gi 158333233:1409149-1410649	Acaryochloris
<i>Anabaena variabilis</i> ATCC 29413	1488	NC_007413	gi 75906225:1002918-1004405	Anabaena
<i>Arthrospira maxima</i> CS-328	1482	NZ_ABYK01000016	gi 209525145:29326-30807	Arthrospira
<i>Crocospaera watsonii</i> WH 8501	1408	NZ_AADV02000003	gi 67921358:105686-107093	Crocospaera
<i>Cyanobium</i> sp. PCC 7001	1481	NZ_ABSE01000018	gi 194271858:58330-59810	Cyanobium
<i>Cyanothece</i> sp. ATCC 51142	1489	NC_010546	gi 172034917:3950932-3952420	Cyanothece 51142
<i>Cyanothece</i> sp. CCY0110	1479	NZ_AAXW01000002	gi 126655026:c195826-194348	Cyanothece 0110
<i>Cyanothece</i> sp. PCC 7424	1483	NZ_ABOY01000032	gi 186903529:26610-28092	Cyanothece 7424
<i>Cyanothece</i> sp. PCC 7425 [†]	N/A	N/A	N/A	N/A
<i>Cyanothece</i> sp. PCC 7822	1483	NZ_ABVE01000002	gi 196255427:394510-395992	Cyanothece 7822
<i>Cyanothece</i> sp. PCC 8801 [†]	N/A	N/A	N/A	N/A
<i>Gloeobacter violaceus</i> PCC 7421	1485	NC_005125	gi 37519569:1571231-1572715	Gloeobacter
<i>Lyngbya</i> sp. PCC 8106	1493	NZ_AAVU01000018	gi 119488018:81300-82792	Lyngbya
<i>Microcoleus chthonoplastes</i> PCC 7420	1488	NZ_ABRS01000062	gi 194018314:60739-62226	Microcoleus

<i>Microcystis aeruginosa</i> NIES-843	1477	NC_010296	gi 166085114:1885814-1887290	Microcystis
<i>Nostoc punctiforme</i> PCC 73102	1410	AF027655	gi 186463002:2021489-2022977	Nostoc 73102
<i>Nostoc</i> sp. PCC 7120	1489	NC_003272	gi 17227497:2375734-2377222	Nostoc 7120
<i>Prochlorococcus marinus</i> str. AS 9601	1465	NC_008816	gi 123967536:323018-324482	Pro 9601
<i>Prochlorococcus marinus</i> str. MIT 9211	1465	NC_009976	gi 159902540:342283-343747	Pro 9211
<i>Prochlorococcus marinus</i> str. MIT 9215	1465	NC_009840	gi 157412338:319913-321377	Pro 9215
<i>Prochlorococcus marinus</i> str. MIT 9301	1465	NC_009091	gi 126695337:322587-324084	Pro 9301
<i>Prochlorococcus marinus</i> str. MIT 9303	1401	NC_008820	gi 124021714:243682-245082	Pro 9303
<i>Prochlorococcus marinus</i> str. MIT 9312	1465	NC_007577	gi 78778385:317580-319044	Pro 9312
<i>Prochlorococcus marinus</i> str. MIT 9313	1465	NC_005071	gi 33862273:208864-210328	Pro 9313
<i>Prochlorococcus marinus</i> str. MIT 9515	1465	NC_008817	gi 123965234:332422-333886	Pro 9515
<i>Prochlorococcus marinus</i> str. NATL1A	1465	NC_008819	gi 124024712:385837-387301	Pro NATL1A
<i>Prochlorococcus marinus</i> str. NATL2A	1394	NC_007335	gi 162958048:370879-372272	Pro NATL2A
<i>Prochlorococcus marinus</i> subsp. <i>marinus</i> str. CCMP1375	1465	NC_005042	gi 33239452:353331-354795	Pro 1375
<i>Prochlorococcus marinus</i> subsp. <i>pastoris</i> str. CCMP1986	1465	NC_005072	gi 33860560:313061-314525	Pro 1986

<i>Synechococcus elongatus</i> PCC 7942	1490	NC_007604	gi 81298811:568589-570078	Sel PCC 7942
<i>Synechococcus</i> sp. CC9311	1477	NC_008319	gi 113952711:529961-531437	Syn CC9311
<i>Synechococcus</i> sp. CC9605	1440	NC_007516	gi 78211558:473828-475267	Syn CC9605
<i>Synechococcus</i> sp. CC9902	1479	NC_007513	gi 78183584:373319-374797	Syn CC9902
<i>Synechococcus</i> sp. JA-2-3B'a(2-13)	1323	NC_007776	gi 86607503:1447574-1448896	Syn JA2
<i>Synechococcus</i> sp. JA-3-3Ab	1324	NC_007775	gi 86604733:2310397-2311720	Syn JA3
<i>Synechococcus</i> sp. PCC 7002	1452	NC_010475	gi 170076636:1461361-1462812	Syn PCC 7002
<i>Synechococcus</i> sp. RCC307	1498	NC_009482	gi 148241099:348737-350234	Syn RCC307
<i>Synechococcus</i> sp. RS9916	1439	NZ_AAUA01000001	gi 116072916:c817919-816481	Syn RS9916
<i>Synechococcus</i> sp. WH 5701	1440	AY172832	gi 87303064:1-2054	Syn WH 5701
<i>Synechococcus</i> sp. WH 7803	1497	NC_009481	gi 148238336:534540-536036	Syn WH 7803
<i>Synechococcus</i> sp. WH 8102	1464	NC_005070	gi 33864539:2083060-2081597	Syn WH 8102
<i>Synechocystis</i> sp. PCC 6803	1489	NC_000911	gi 16329170:3325053-3326541	Synechocystis
<i>Thermosynechococcus elongatus</i> BP-1	1491	NC_004113	gi 22297544:2335243-2336733	Thermosynechococcus
<i>Trichodesmium erythraeum</i> IMS101	1482	NC_008312	gi 113473942:3137164-3138645	Trichodesmium
<i>Rhodobacter sphaeroides</i> ATCC 17029	1460	NC_009049	gi 126460778:87772-89231	Rhodobacter

* refers to the concatenated 16S rRNA and 23S rRNA sequences

† either absent in the GenBank or not used because of absence of the respective 23S rRNA

Table S3 – List of the 23S rRNA sequences used in the study.

Species and strain	Length, bp	GenBank nucleotide sequence accession number	GI/EMB List
<i>Acaryochloris marina</i> MBIC11017	2880	NC_009925	gi 158333233:5638205-5641084
<i>Anabaena variabilis</i> ATCC 29413	2830	NC_007413	gi 75906225:3896066-3898895
<i>Arthrospira maxima</i> CS-328	2882	NZ_ABYK01000016.1	gi 209494330:31287-34168
<i>Crocospheara watsonii</i> WH 8501	2881	NZ_AADV02000003.1	gi 67856470:107505-110384
<i>Cyanobium</i> sp. PCC 7001	2886	NZ_ABSE01000005.1	gi 194271845:94-2979
<i>Cyanothece</i> sp. ATCC 51142	2875	NC_010546	gi 172034917:3952751-3955625
<i>Cyanothece</i> sp. CCY 0110	2862	NZ_AAXW01000002	gi 126655026:c193919-191058
<i>Cyanothece</i> sp. PCC 7424	2886	NZ_ABOY01000032.1	gi 186689375:28427-31312
<i>Cyanothece</i> sp. PCC 7425*	N/A	N/A	N/A
<i>Cyanothece</i> sp. PCC 7822	2882	NZ_ABVE01000004	gi 196255427:c391299-394180
<i>Cyanothece</i> sp. PCC 8801*	N/A	N/A	N/A
<i>Gloeobacter violaceus</i> PCC 7421	2781	NC_005125	gi 37519569:c1570772-1567992
<i>Lyngbya</i> sp. PCC 8106	2878	NZ_AAVU01000018.1	gi 119488018:83267-86144
<i>Microcoleus chthonoplastes</i> PCC 7420	2884	NZ_ABRS01000062.1	gi 194018314:62741-65624

<i>Microcystis aeruginosa</i> NIES-843	2878	NC_010296	gi 166362741:1887656-1890533
<i>Nostoc punctiforme</i> PCC 73102	2887	NC_010628	gi 186680550:2023598-2026484
<i>Nostoc</i> sp. PCC 7120	2828	NC_003272	gi 17227497:2377736-2382034
<i>Prochlorococcus marinus</i> str. AS 9601	2874	NC_008816	gi 123967536:325023-327896
<i>Prochlorococcus marinus</i> str. MIT 9211	2876	NC_009976	gi 159902540:344441-347316
<i>Prochlorococcus marinus</i> str. MIT 9215	2874	NC_009840	gi 157412338:321917-324790
<i>Prochlorococcus marinus</i> str. MIT 9301	2876	NC_009091	gi 126695337:324618-327493
<i>Prochlorococcus marinus</i> str. MIT 9303	2873	NC_008820	gi 124021714:245970-248842
<i>Prochlorococcus marinus</i> str. MIT 9312	2874	NC_007577	gi 78778385:319585-322458
<i>Prochlorococcus marinus</i> str. MIT 9313	2876	NC_005071	gi 33862273:211158-214033
<i>Prochlorococcus marinus</i> str. MIT 9515	2874	NC_008817	gi 123965234:334434-337307
<i>Prochlorococcus marinus</i> str. NATL1A	2874	NC_008819	gi 124024712:387935-390808
<i>Prochlorococcus marinus</i> str. NATL2A	2875	NC_007335	gi 162958048:372962-375836
<i>Prochlorococcus marinus</i> subsp. <i>marinus</i> str. CCMP1375	2873	NC_005042	gi 33239452:355462-358334
<i>Prochlorococcus marinus</i> subsp. <i>pastoris</i>	2875	NC_005072	gi 33860560:315074-317948

str. CCMP1986			
<i>Synechococcus elongatus</i> PCC 7942	2878	CP000100.1	gi 81298811:570624-573501
<i>Synechococcus</i> sp. CC9311	2864	NC_008319	gi 113952711:532200-535063
<i>Synechococcus</i> sp. CC9605	2866	NC_007516	gi 78211558:476063-478928
<i>Synechococcus</i> sp. CC9902	2866	NC_007513	gi 78183584:375574-378439
<i>Synechococcus</i> sp. JA-2-3B'a(2-13)	2808	NC_007776	gi 86607503:1449625-1452432
<i>Synechococcus</i> sp. JA-3-3Ab	2807	NC_007775	gi 86604733:c1109571-1106765
<i>Synechococcus</i> sp. PCC 7002	2811	NC_010475	gi 170076636:1463345-1466155
<i>Synechococcus</i> sp. RCC307	2866	NC_009482	gi 148241099:350832-353697
<i>Synechococcus</i> sp. RS9916	2865	NZ_AAUA01000004.1	gi 116072916:c812870-815734
<i>Synechococcus</i> sp. WH 5701	2884	NZ_AANO01000001.1	gi 87282231:2913-5796
<i>Synechococcus</i> sp. WH 7803	2866	NC_009481	gi 148238336:536808-539673
<i>Synechococcus</i> sp. WH 8102	2865	NC_005070	gi 33864539:c1870985-1873849
<i>Synechocystis</i> sp. PCC 6803	2883	NC_000911	gi 16329170:c2451721-2448839
<i>Thermosynechococcus elongatus</i> BP-1	2871	NC_004113	gi 22297544:c2334822-2331207
<i>Trichodesmium erythraeum</i> IMS101	2881	NC_008312	gi 113473942:3139115-3141995
<i>Rhodobacter sphaeroides</i> ATCC 17029	2918	NC_009049	gi 126460778:89899-92816

* either absent in the GenBank or not used because of absence of the respective 16S rRNA

Figure S1

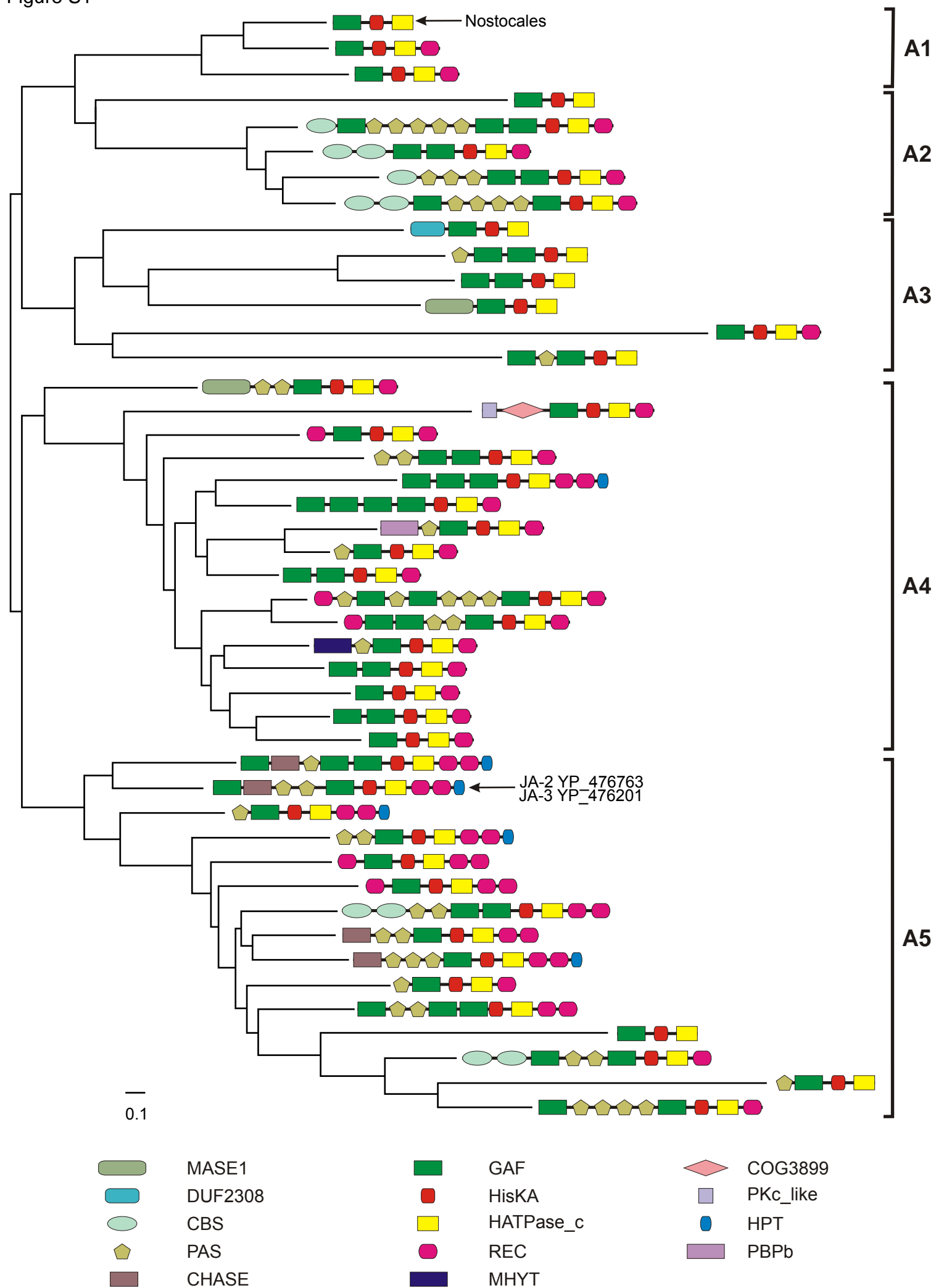


Figure S2

		*	20	*	40	*	60	*	80	*		
Sel_Cika	:	-----MLAPSSNCSLASQRITPEG									:	19
Ac_YP_001518724	:	-----MMPVSAPEIFEHTLPQHV									:	18
An_YP_321687	:	-----MLSSPDLSFSRNLPLVV									:	17
Ar_ZP_03273921	:	-----									:	-
Cr_ZP_00513747	:	-----MNPVSKSILRRTLPISI									:	17
Cy0110_ZP_01728511	:	-----MNPVSKSILRRTLPISI									:	17
Cy7424_ZP_02976063	:	-----MNSPQMFVFRRLVSQQT									:	17
Cy7425_ZP_03139397	:	-----MLQMOTVTIDQATPAPQ									:	17
Cy7822_ZP_03155846	:	-----MDSSQVFVFRRKVSQQT									:	17
Cy8801_ZP_02940946	:	-----MTRFQDSILRRTLSMGT									:	17
CyATCC_YP_001806165	:	-----MSNMNPVSKSILRRTLPISI									:	21
Ly_ZP_01622671	:	-----									:	-
Mc_YP_002622050	:	MAILGNPRFSRTLPLAMFERLQELLQQMQGKIGREAVLLRSQDLPIVSVGLDQQVQRFTLLVSPGFRAILLIATPIESQGEAEANSHSSLYQVGLTF									:	96
Mi_YP_001659385	:	-----MNASPAYVCCQVLSSTH									:	17
No_NP_485728	:	-----MLSSPDLSFSRNLPLVV									:	17
Np_ZP_00108622	:	-----MLSSPDLSFSRTLPIVV									:	17
Scy_NP_441517	:	-----MLPAFSPIFRRLIPAVT									:	17
Syn_YP_001734160	:	-----MILNAITPDIAWTAPWNV									:	18
Th_NP_681689	:	-----MPQPIFDRIIPAFI									:	14
Tr_YP_721187	:	-----MAISYKIIIGDMPSFPPAKPDILPLNI									:	27



		100	*	120	*	140	*	160	*	180	*		
Sel_Cika	:	FAQTQSALQDFVA--TLPQAFYWDSRSLHHTLRTQTGDCAIAIAAGFQLLLIGRTA-----										:	73
Ac_YP_001518724	:	YPQISLSLADMVR--SVPGVAKLLTSVEVGAMVPETQRFTVLTSKSFSAIVYEAKE-----										:	72
An_YP_321687	:	FNQIGELLQQMAQ-----SVETSTLLLTESLLSRVFPMEWHSQRFTLVVSEGFSALLLGNWEQQELPESENWILAMPSPILNQGIAGDRETQK										:	105
Ar_ZP_03273921	:	-----										:	-
Cr_ZP_00513747	:	FEKICSLWRELME-----IEGSSAILLNDISIYPEITDISEILGAKKFYLLSSQ-----										:	67
Cy0110_ZP_01728511	:	FEEICSLWRELVE-----IEDSTAILLNDISIISSEIVNFSEILGARTFYLLSS-----										:	66
Cy7424_ZP_02976063	:	LLAISHILLEQMSA----QFGQEASLWTEQTLFQDHQDFPDNPSEIDYWRLLISPO-----										:	68
Cy7425_ZP_03139397	:	WEQIVAVLLQAVP-----PLASPLVILTEQEVSGYPPTVDRLVVVISAAPK-----										:	63
Cy7822_ZP_03155846	:	LSAICRLLEQISA----HIGQEAILLTQETLFPEQQRGKNPWGTECLOLLIAPQ-----										:	67
Cy8801_ZP_02940946	:	FEEICSLWRQIVE-----MGETESLLITQETILAEIPHPQAPLTREQFSIFLSPE-----										:	67
CyATCC_YP_001806165	:	FEEICSLWRELVE-----IEDSTAILLNDISIISSEIVNVPEILGAKTFYLLSS-----										:	70
Ly_ZP_01622671	:	-----										:	-
Mc_YP_002622050	:	ESDAIAGFLDGFTRQFRTPQPIILHVFKLVQTYPQANQAAIQSEFTLLLLMDILTANP-----										:	152
Mi_YP_001659385	:	LEQIRSWWGQIAI----QVTAPRISLSERELPPQTGEIYQLLLSTDLOALLASP-----										:	68
No_NP_485728	:	FNQIGELLQQMAQ-----SVETSTLLLTESLLSRVLVPMEWHGQRFTLVVSEGFSALLLGNKGQQEVPESSESVWLAMPSPILNQSATGDRETQK										:	105
Np_ZP_00108622	:	FNQIGELLEQMAQ-----TVGSATLVLTEAVLARISIPVEWQRQRFTLVVSGQFSALLLGNFEEQGSGRGAGEQGGSTSL-----RDAARTTT										:	100
Scy_NP_441517	:	FERILRFWRTLAQ--QTGDGVQCFVGDLPSSSLKPPPGPSVLEAEVDHRFALLVSPG-----										:	71
Syn_YP_001734160	:	LQQICHLWRRFGIGIDPQDRLYISGAQLQEAGLALDYDFHLLIETDFTAILWAKA-----										:	74
Th_NP_681689	:	YERLATVLLAQAS-----RRGATVLTREEVIASTDAPFLIVVAESFALLQAE-----										:	62
Tr_YP_721187	:	LEPISQLLQKTVQ--GLDSIIITEDILATTQSLIFPGVKFTAVISKKFNGLLWGKL-----IKDLSNPKSQNR-----EQQEK										:	99



	200	*	220	*	240	*	260	*	280	
Sel_Cika	: -----AEYCQPHPLSEPHHVS		QFGADSI		QRYCQATNLPVEY-----		QPALAAQLGDL		SLNPDLLISQFSNLLIAAIAADRA	: 143
Ac_YP_001518724	: -----AQLK		TFEPMA		TTFLER		YQHCPA-GSPLQQR		I-----TEVRERQSRKNNDAKLQ	: 138
An_YP_321687	: QEIYSASSIPQAAQSPLTL		TELNTR		TFNSEA		ASFLAK		IRDLFGLD-DSPTHQHL-----	: 190
Ar_ZP_03273921	: -----									: -
Cr_ZP_00513747	: -----LCALLQGT		LNDSLSYQVI		ITCDPQA		AEFTHK		IEQHLDN-SSPWRKRL-----	: 146
Cy0110_ZP_01728511	: -----QLCALLQGT		LNASSLSYQVT		ITCEPQA		AEFTHK		IQQNLDN-DSPWHDR	: 146
Cy7424_ZP_02976063	: -----LNVLLLGK		FIVHQSLSYQIS		ITWDSQT		IQDLLTQ		IAQERTL-----SQHLPL	: 141
Cy7425_ZP_03139397	: -----AVLIELKTE		PTYQLQ		IVFSYQD		IKNFLQ		QITPRYPT-DQPLLEWACL	: 148
Cy7822_ZP_03155846	: -----MSILLGNF		SVPDSLSYQIS		ITWDSQT		IQDLLTQ		ITPSVTLPQE-LEQYLTQ-----	: 150
Cy8801_ZP_02940946	: -----FSALLRG		TFDTIALCYQV		MSWDGEA		ADFDY		IKQHLPP-TSVLRDHL-----	: 146
CyATCC_YP_001806165	: -----QLCALLQGT		LNASSLSYQVT		ITCEPQA		AEFTHK		IQQLHDP-ASPWHDR	: 150
Ly_ZP_01622671	: -----									: -
Mc_YP_002622050	: -----				ISNESIHPYCEV		GQP-----			: 167
Mi_YP_001659385	: -----QGHNFQY		EVRI		SFEANT		IKNFLQ		GSVKSLH-NPKIQQGY-----	: 138
No_NP_485728	: PEIYSTSSIPQAAQP		PTLTLTELNTR		TFNSEA		ASFLAK		IRDLFGLD-DSYTHQHL-----	: 190
Np_ZP_00108622	: LSDQGAGGQEI		IHSLLSTQNSALNAR		TFNLEA		ASFLHE		IRDLFEC-DSYTHQNL-----	: 185
Scy_NP_441517	: -----QWALLEGE		QISPHHYAVS		ITFAQGI		IEDFIQ		KQNLPPVA-----EAMPHRPET	: 142
Syn_YP_001734160	: -----IDPAAVTP		DTLSAPGEQYE		ISMVEDPQA		IAPLIAT		ILGSQVD--FP--QL-----	: 151
Th_NP_681689	: -----PVPQMSTY		RVAILTNPR		ARFLRK		IRSQVPV-----		NRRPLIRAVLQQLSPLNAKEQMLPADLAI	: 127
Tr_YP_721187	: YSLQSNSFSFPFNTKNL		GSLKLNVG		ISFDPEE		IKFFLHQ		INRLEA-DKTKLFI-----	: 187


	*	300	*	320	*	340	*	360	*	380	
Sel_Cika	: PLAAQ-----							YPAVSVCQPI		EQALHWQEE--QDRLLISQVSAQIRLSID	: 184
Ac_YP_001518724	: AE-----							AVPLA--		PEVTILHPTAN--QERLISHQIISQLSQGV	: 172
An_YP_321687	: EEVTASASVNN-----							SEEVYACQ		AVEDALKKQIS--QERLLINQVTTQIRKSID	: 237
Ar_ZP_03273921	: -----							MDDSINKQ		LE--RERLLINQVTTQIRQSID	: 27
Cr_ZP_00513747	: ETSDVM-----							YEGLR		TCEPVEEALRQQVA--QERLLINQVTTQIRQSID	: 188
Cy0110_ZP_01728511	: ETSEVM-----							YEGLR		VCEPVEDALRQQIA--QERLLINQVTTQIRQSID	: 188
Cy7424_ZP_02976063	: ETPASL-----							SPPTS		SLYHPIEEALRQQVE--QERLLINQVTTQIRQSID	: 183
Cy7425_ZP_03139397	: -----							TTVSL		NSPLSPWQQQLITOMTTQIWQNL	: 177
Cy7822_ZP_03155846	: MK-----							CIHAS		VCPVEEALRQQVE--QERLLINQVTTQIRQSID	: 188
Cy8801_ZP_02940946	: ADSQEI-----							YTTAM		VCKPVEDALRQQIA--QERLLINQVTTQIRQSIN	: 188
CyATCC_YP_001806165	: ETSEVM-----							YEGLR		VCEPVEDALRQQVA--QERLLINQVTTQIRQSID	: 192
Ly_ZP_01622671	: -----							MDKAL		DQQL--QERLLINQVTTQIRQSID	: 27
Mc_YP_002622050	: -----							IEAAL		RQQVE--QERLVHQTTHIRQSID	: 194
Mi_YP_001659385	: DRDC-----							SRIFS		VCPVEEALRQQIA--QERLLINQVTTQIRQSID	: 178
No_NP_485728	: ESVTTSASVNN-----							SEEVY		SCQAVEDALKKQIS--QERLLINQVTTQIRKSID	: 237
Np_ZP_00108622	: QEIIAPPSPT-----							APAVY		ICQPVEDALKKQIS--QERLLINQVTTQIRKSID	: 231
Scy_NP_441517	: SFS-----							PEPSL		QDSLQASQVLLISQVIAQIRQSID	: 174
Syn_YP_001734160	: QETAP-----							DEVV		VPNQPLNMLLSQRIE--QEKILNNVTHRIYQND	: 192
Th_NP_681689	: VLGEET-----							TAQCQ		SCQPVTAALNERQA--QERLLINQVTTQIRQSID	: 169
Tr_YP_721187	: IDSIFRMTSNQKPREKKNY		TELSKINFH		SKSTTNT		EFITETNLLQNKLLASDSVMNDQ		EYPLVCKPIEIALNQKLE--QERLLINQVTTQIRQSID	: 281	


		*	400	*	420	*	440	*	460	*	480																																																																																						
Sel_Cika	:	L	S	E	I	L	T	T	I	R	E	I	R	Q	L	N	A	D	R	A	I	I	Y	Q	F	K	P	Q	C	L	D	A	G	L	D	Q	R	W	P	L	Y	I	P	-----	S	Q	S	Y	I	T	T	E	D	R	R	N	E	A	L	S	V	I	D	P	L	V	Q	P	:	253																											
Ac_YP_001518724	:	L	P	A	L	L	S	M	V	E	E	L	R	H	L	L	Q	A	D	R	L	L	I	Y	-	Q	L	S	P	T	A	P	S	G	A	E	S	H	H	P	V	S	A	-----	Q	S	G	S	V	V	H	E	A	L	A	S	S	E	L	P	S	V	L	S	--	D	Q	E	:	237																											
An_YP_321687	:	L	P	V	I	M	A	T	A	I	A	Q	V	R	E	F	L	E	D	R	L	V	I	Y	K	F	E	G	S	R	V	N	T	Q	N	T	R	P	P	R	L	E	D	W	Q	N	-----	Y	G	G	C	I	V	Y	E	A	R	A	T	D	I	I	P	S	V	L	D	-	Y	Q	E	K	:	307																							
Ar_ZP_03273921	:	L	S	V	I	L	S	T	A	V	Q	Q	L	R	E	F	L	Q	V	D	R	L	V	I	Y	Q	L	N	S	Y	Q	V	T	G	D	H	T	K	D	L	S	P	D	V	D	E	P	D	L	N	T	E	I	P	V	P	Q	A	D	R	H	D	Q	-----	I	H	G	S	I	T	Y	E	A	K	S	S	N	A	I	A	S	L	L	D	W	E	E	A	D	:	115						
Cr_ZP_00513747	:	L	S	V	I	L	E	T	A	V	R	E	L	R	S	F	L	E	V	D	R	L	F	I	Y	K	F	R	E	K	Y	L	S	T	S	E	S	E	K	I	Q	N	-----	L	G	G	L	V	T	Y	E	S	R	V	S	Q	S	I	P	S	L	L	N	V	V	A	E	D	:	254																											
Cy0110_ZP_01728511	:	L	S	V	I	L	Q	T	A	V	R	E	L	R	S	F	L	Q	V	D	R	L	V	I	Y	E	F	K	E	F	K	E	K	I	I	S	N	S	S	S	R	N	S	R	Q	-----	T	G	G	L	V	T	Y	E	S	R	V	S	Q	S	I	P	S	L	L	N	V	V	A	E	D	:	257																								
Cy7424_ZP_02976063	:	L	S	Q	V	L	K	T	A	V	T	E	V	R	N	F	L	G	V	D	R	L	L	I	Y	Q	F	K	F	N	S	S	V	T	D	S	T	P	H	S	L	T	-----	R	K	A	T	I	T	Y	E	A	R	A	S	Q	L	I	P	S	M	L	H	L	S	P	E	E	:	249																											
Cy7425_ZP_03139397	:	L	S	L	I	L	Q	T	T	L	E	Q	V	Q	R	V	L	E	V	D	R	L	L	I	Y	Q	L	Q	-----	T	H	H	Q	I	T	H	E	A	L	A	A	H	A	P	P	S	L	L	G	-	K	G	D	P	:	228																																									
Cy7822_ZP_03155846	:	L	P	V	V	L	E	T	A	V	T	E	V	R	N	F	L	Q	V	D	R	L	L	I	Y	Q	F	S	S	H	P	S	E	T	E	T	K	Q	T	F	P	N	-----	G	W	C	K	I	T	Y	E	A	R	A	S	Q	L	V	P	S	L	L	N	M	I	P	E	D	:	254																											
Cy8801_ZP_02940946	:	L	S	A	I	L	K	T	A	V	R	E	V	R	S	F	L	Q	V	D	R	L	V	I	Y	E	F	G	Q	G	T	P	P	N	S	E	Y	S	Q	P	F	T	-----	S	W	G	C	V	T	Y	E	S	K	V	S	N	S	I	P	S	L	L	N	V	V	A	E	D	:	254																											
CyATCC_YP_001806165	:	L	S	V	I	L	E	T	A	V	R	E	L	R	S	F	L	Q	V	D	R	L	V	I	Y	E	F	R	E	K	T	I	S	N	Q	S	E	N	A	Q	K	-----	T	S	C	L	V	T	Y	E	S	R	V	S	Q	S	I	P	S	L	L	N	L	V	A	E	D	:	258																												
Ly_ZP_01622671	:	L	S	I	I	L	S	T	T	I	V	E	Q	L	Q	F	L	Q	V	D	R	L	L	I	Y	Q	F	D	F	Q	S	I	P	S	R	E	C	S	L	K	P	R	L	D	L	E	E	Q	N	R	D	T	R	L	T	P	P	L	P	K	D	S	S	R	S	K	P	S	E	K	G	Y	K	C	V	T	Y	E	A	R	S	S	E	N	I	L	S	V	L	H	L	N	E	T	D	:	123
Mc_YP_002622050	:	L	P	V	I	L	S	T	A	V	D	E	V	R	H	F	L	N	V	D	R	L	L	I	Y	Q	F	D	G	D	L	V	N	L	R	S	E	G	D	K	I	E	V	D	N	S	T	D	V	T	D	E	F	E	N	A	Q	F	L	T	S	S	P	L	S	Q	Q	-----	D	E	C	C	V	T	Y	E	A	R	S	S	E	I	R	S	V	L	N	W	R	E	E	K	:	285			
Mi_YP_001659385	:	L	P	V	I	L	E	T	V	V	R	E	A	R	E	F	L	Q	V	D	R	L	L	I	Y	Q	F	F	P	S	T	S	E	-----	V	K	G	I	T	S	E	R	I	S	E	Q	I	N	S	V	L	N	L	T	P	E	D	:	235																																						
No_NP_485728	:	L	P	V	I	M	A	T	A	I	A	Q	V	R	E	F	L	E	D	R	L	V	I	Y	K	F	E	G	S	G	V	N	T	Q	N	T	R	S	P	R	L	Q	D	W	Q	N	-----	Y	G	G	C	I	V	Y	E	A	R	A	T	D	I	I	P	S	V	L	D	-	Y	Q	E	Q	:	307																							
Np_ZP_00108622	:	L	P	V	I	M	A	T	A	I	T	Q	V	R	E	F	L	E	D	R	L	V	I	Y	K	F	E	A	S	K	V	K	T	Q	E	Y	Q	S	S	T	D	E	D	N	G	K	G	S	T	S	I	S	V	N	N	Q	S	L	L	E	D	Y	Q	Q	-----	H	R	G	Y	I	V	Y	E	A	R	A	T	D	A	I	T	C	V	L	N	-	Y	T	E	K	:	319					
Scy_NP_441517	:	L	S	E	I	L	N	N	A	V	T	A	Q	K	F	L	F	V	D	R	L	V	I	Y	Q	F	H	Y	S	Q	P	S	L	T	P	L	E	E	N	Q	I	P	A	P	R	P	R	Q	-----	Q	Y	C	E	V	T	Y	E	A	R	R	S	P	E	I	D	T	M	L	G	I	M	T	E	N	:	247																					
Syn_YP_001734160	:	L	M	V	T	V	R	M	A	L	E	Q	A	Q	R	L	L	R	V	D	R	L	L	V	Y	Q	L	D	L	P	T	A	K	P	E	K	-----	F	V	N	R	V	T	F	E	V	V	S	S	D	K	V	T	S	L	L	H	-	F	D	Q	D	:	251																																	
Th_NP_681689	:	L	P	E	L	L	K	I	A	V	D	R	I	R	E	F	L	D	V	D	R	L	L	V	G	F	A	Q	T	E	G	E	-----	L	R	G	Q	I	T	Y	E	S	C	R	N	S	E	I	P	S	V	L	G	I	W	D	D	C	:	226																																					
Tr_YP_721187	:	L	P	E	I	L	S	T	A	V	K	E	V	R	K	F	I	Q	V	D	R	L	L	I	Y	E	L	N	N	P	L	L	L	T	D	N	I	T	T	V	N	E	-----	Q	K	G	S	V	K	Y	E	S	L	V	S	N	K	I	S	S	L	L	G	L	S	E	R	K	:	347																											

		*	500	*	520	*	540	*	560	*																																																																																									
Sel_Cika	:	-	G	L	L	I	-	T	T	E	E	W	Q	R	F	Q	Q	E	T	L	L	D	S	V	G	F	Y	K	E	R	L	P	E	Q	Y	S	F	Y	E	R	V	Q	V	R	S	V	C	K	I	P	I	L	V	Q	G	R	I	W	G	L	L	V	A	H	Q	C	Q	Q	D	H	R	W	Q	P	R	E	R	D	I	L	Q	H	L	A	E	H	L	S	I	A	I	Y	Q	A	Q	:	347		
Ac_YP_001518724	:	Q	L	W	G	I	A	S	S	E	S	L	L	K	Y	Q	Q	Y	T	H	A	T	H	D	V	L	Q	N	D	S	V	S	E	P	I	A	K	V	I	D	S	I	Q	V	R	A	L	L	V	T	P	I	L	V	K	S	E	L	W	G	L	L	I	A	H	Q	C	L	H	P	R	Q	W	Q	V	Q	E	Q	D	L	I	K	R	M	A	E	H	V	A	I	A	I	H	Q	A	:	333		
An_YP_321687	:	T	C	H	S	R	H	S	-	Q	C	W	E	K	Y	R	Q	F	T	L	V	I	D	D	I	E	Q	A	V	A	L	E	E	C	L	V	N	F	L	R	E	N	Q	V	R	A	K	I	A	A	P	I	I	F	E	D	K	L	W	G	L	L	I	A	H	Q	C	H	S	P	R	Q	W	N	D	S	D	K	N	L	L	I	S	I	A	E	Q	L	A	I	A	I	H	Q	S	:	402		
Ar_ZP_03273921	:	R	C	H	A	Q	V	P	-	E	Y	Q	Q	Y	Q	K	S	I	I	H	C	I	P	N	I	K	L	Y	Q	S	S	P	E	L	Y	S	I	L	Q	Q	K	Q	V	T	A	Q	L	V	A	P	I	L	V	Q	K	L	W	G	I	L	I	A	H	Q	C	F	S	D	R	Q	W	K	E	S	E	Q	K	F	I	S	K	I	T	O	H	L	S	I	A	I	Y	Q	A	Q	:	210			
Cr_ZP_00513747	:	V	C	H	A	T	I	P	-	E	Y	Q	H	K	Y	R	Q	E	T	I	V	A	T	E	D	V	E	R	Q	S	A	S	L	C	L	S	Q	F	L	E	K	Y	W	I	L	S	K	L	I	A	P	I	I	V	E	G	E	L	W	G	L	L	I	A	H	Q	C	F	K	K	R	Q	W	L	E	S	E	K	A	F	L	G	Q	I	G	E	H	L	A	V	A	I	Y	Q	A	Q	:	349	
Cy0110_ZP_01728511	:	T	C	H	T	N	I	P	-	Q	Y	K	Y	K	Y	R	Q	C	Q	I	V	S	I	D	D	V	E	M	R	Y	A	A	S	L	C	L	S	R	F	L	E	Q	Y	W	V	L	S	K	L	I	A	P	I	V	V	N	G	K	L	W	G	L	L	I	A	H	Q	C	F	K	K	R	Q	W	L	E	S	E	K	A	F	L	G	Q	I	G	E	H	L	A	V	A	I	Y	Q	A	Q	:	352
Cy7424_ZP_02976063	:	D	C	F	V	C	I	P	-	F	Y	Q	E	K	Y	R	Q	C	T	V	V	A	V	E	N	A	E	T	E	Y	S	S	S	F	C	L	A	E	L	R	Q	N	H	V	K	S	L	L	I	A	P	I	I	V	D	E	Q	L	W	G	L	L	I	A	H	Q	C	F	K	I	R	Q	W	L	E	S	E	K	G	F	L	G	H	I	G	E	H	L	A	I	A	I	Q	Q	A	Q	:	344	
Cy7425_ZP_03139397	:	T	C	L	H	S	R	T	-	A	C	Y	Q	R	Y	L	Q	E	I	Q	V	V	S	D	V	E	A	T	Y	A	D	T	P	C	L	L	A	L	Q	Q	S	Q	V	R	A	K	L	I</																																																			

	580		*	600	*		620	*	640	*	660	*				
Sel_Cika	: IYGO-----IQDQTQTL	ENRVL	ERTQEL	LDALALAA	ANAAG	GEFLAT	MSHEL	RTP	LT	CVIGMS	STLL	RWAFG	PLT-----ERQREYIKAI	IHD : 430		
Ac_YP_001518724	: IQDQ-----VROHKQ	TLDKQ	VEORTQ	ELHAALLA	ASSNQ	AKSDFL	ATMSHEL	RTP	LT	CVIGMS	SATLL	RWSLG	PLT-----DKQRSCLQ	IHD : 416		
An_YP_321687	: IMRSLQDA	AHRLTQ	EKHTLE	QRVIER	TMAIRD	ALLAAE	AASRLR	SEFLAT	ISHELL	TPL	TYVIG	MSSTLL	RWPLG	ELS-----QRORDYLQ	IHD : 492	
Ar_ZP_03273921	: IYAO-----VQQQKQ	TLERRV	KERTQ	ALHDALL	AAQCAN	RAKSE	FLATMS	HEL	RTP	LTH	VIGIS	STLL	RLYSQ	PDNFPALPLEKQ	KHYLQTIHD : 299	
Cr_ZP_00513747	: IYAO-----VQEQKN	TFEQRV	IESTQ	ALRDTL	IASQA	ANHSK	SEFLGN	MSHEL	RTP	LT	CIIGL	SGTLL	HWSQ	EGLN---LPVDKQ	KYLDTIQN : 435	
Cy0110_ZP_01728511	: IYAO-----VQEQKN	NNEFQR	VIERTQ	ELRDTL	VASQA	ANHSK	SEFLGN	MSHEL	RTP	LT	CIIGL	SGTLL	HWSQ	ESSN---LPIDKQ	KYLDTIQN : 438	
Cy7424_ZP_02976063	: IYQO-----VQQQKQ	NNEFQR	VIECTE	EELKNT	IAVAQ	SAHLSK	SEFLSN	ISHELL	TPL	TCVIG	L	AGTLL	HWSG	EQSS---LSPEKQ	KYIESIQ : 430	
Cy7425_ZP_03139397	: IQSQ-----LQQQKQ	TLEQRV	IER	TQALYD	ALLSTH	SAHRAK	SDFLAT	INHEL	RTP	LT	TCVIG	MSATLL	RWSLG	PLN---DKQRSY	LKTIHD : 406	
Cy7822_ZP_03155846	: IYKO-----VQEQKQ	FFEKRV	FECTE	ELRSSI	IAVAQ	SAHLSK	SEFLSN	ISHELL	TPL	TCVIG	L	AGTLL	HWSG	EQSS---LSPEKQ	KYIESIQ : 435	
Cy8801_ZP_02940946	: IYAO-----VQEQKT	TFEQRV	IER	TQELRD	TLIAA	QSANHSK	SEFLSN	MSHEL	RTP	LT	CIIGL	SGTLL	HWSA	QSKS---LPLQKQ	QYLDTIQ : 435	
CyATCC_YP_001806165	: IYAO-----VQEQKN	TFEQRV	IER	TQELRD	TLVAS	QAANHSK	SEFLGN	MSHEL	RTP	LT	CIIGL	SGTLL	HWSQ	ESTT---LPLDKQ	KYLDTIQN : 439	
Ly_ZP_01622671	: LHTQ-----LQQQKY	TLERRV	TERTQ	ALHDALL	AAQSAN	RAKSE	FLATMS	HEL	RTP	LT	TCVIG	ISATLL	RLYS	NGAGIKQISREKQ	QYELHKIHD : 308	
Mc_YP_002622050	: IYAO-----LHQKSS	LEQRV	IER	TQELRD	ALQAT	QAANHA	KSEFLA	AMSHEL	RTP	LT	TCVIG	L	SATLL	RWSLG	EGGSKTVSIEKQ	RRYLQTIQ : 469
Mi_YP_001659385	: IYSQ-----VQEQKD	IFERRV	IER	TQELQD	TLMAA	EAANL	CKSEFL	NNISHELL	TPL	TCIIG	L	STTL	QKCSA	ANNF---LPPDKQ	KHYLKVIQ : 416	
No_NP_485728	: IMRSLQDS	AQRILTQ	EKQTL	EQRVIER	TMAIRD	ALLAAE	AASRLR	SEFLAT	ISHELL	TPL	TYVIG	MSSTLL	RWPLG	ELS-----QRORDYLQ	IHD : 492	
Np_ZP_00108622	: IMQT-----LTQEKQ	TLEQRV	IER	TIGLRD	ALLAAE	AANRLR	SEFLAT	ISHELL	TPL	TYVIG	MSSTLL	RWPLG	ELS-----QRORDYLQ	IHD : 497		
Scy_NP_441517	: IYSE-----VQKQKN	NNEFQR	VIERTK	ELRDTL	MAAQA	ANLLK	SQFINN	ISHEL	RTP	LT	SIIGL	SATLL	RWFDP	HAS---LPPAKQ	QYLLNIQ : 428	
Syn_YP_001734160	: SYQO-----LQEQKT	SLETLA	QRARE	LEDALL	SACVA	SQSKKE	FIHIMS	HELL	TPL	TSIIG	L	SNTLS	YWTAD	DNP-KKLSPE	KQRYLQTIH : 434	
Th_NP_681689	: IYSE-----LQQQKA	TLEQRV	NERTQ	ALREAL	SAMTA	AHRIK	NDFLAT	MSHEL	RTP	LT	TCVIG	V	SATLL	RWPLG	PLT-----AKQREY	LEIIEH : 405
Tr_YP_721187	: IYGE-----LQKQKQ	TLEKRV	VERIQD	LYDAVQ	SAE	SANRAK	SEFLAT	MSHEL	RTP	LT	TCVIG	ISSTLL	QWSY	GNRGAK	KMPIQKORDYL	KIHD : 530

	680		*	700	*	720	*	740		*	760																				
Sel_Cika	: SGEHLLE	LINDIL	DL	SLQIB	AGKAALQ	VRPFS	LSRLAT	QTINTL	QEKAR	LGEI	QIMLD	LQINN	RV	V	ERAD	PKRLRQ	ILINLL	SNAVK	FTEP	QGT	VF	: 526									
Ac_YP_001518724	: SGEHLLE	LINDIL	DFSH	VQSG	KATLN	VSD	SLTTL	VQQLQ	VMRD	KADAHQ	VQ	LKAN	LKIP	PERD	RE	FIAD	LRRV	QOIL	ISLTD	NAIK	FT	PMGGK	VT : 512								
An_YP_321687	: SGEHL	DMIND	IL	DL	SLQIB	AGKTV	LN	ABFSL	IKVA	ENTIES	LVEK	AL	SEKVN	LKLD	LQID	PRR	R	FTADA	ARIEQ	ILWN	LL	TNAIK	FT	PEG	GNVT : 588						
Ar_ZP_03273921	: SGEN	LALIND	IL	DL	SLQV	B	SGKT	VLKF	SEFSL	QKLAQ	QC	LHTLR	DF	AKEG	VNVN	LQIK	VDQD--	LFRAD	YRRV	ROIL	LNLL	SNAIK	FT	TEK	GGRVT : 393						
Cr_ZP_00513747	: SKHLL	DMINE	ILEY	SNLQ	SGKY	VLSV	RYFSL	TKVAK	NVMQ	RVSD	ESEH	RRIN	LELD	LQIE	PQK	S	SADP	ERVQ	OILY	HL	LN	NAL	KFT	SKN	GKVT : 531						
Cy0110_ZP_01728511	: SKHLL	DMINE	ILEY	SNLQ	SGKY	VLA	REFSL	MKVAK	NVTQ	RVSD	EAEH	RRIN	LELD	LQINA	Q	S	FYAD	PERV	QOIL	YHL	LN	NAL	KFT	SENG	TVT : 534						
Cy7424_ZP_02976063	: NGKKL	MDLIND	IL	DYSS	ITS	ENYQ	LR	REFS	LSY	VASAV	M	GDFQ	EEA	QKKS	IELV	LD	FQV	KKEE	IKFY	AD	PDRV	QOIL	ISHL	M	NAIK	FT	PE	GGRVT : 526			
Cy7425_ZP_03139397	: SGEHLLE	LINDIL	DL	SLQV	EAG	AVLS	SEFSL	NQ	CRQA	RV	FREKA	QEA	EVEL	KLES	LIP	SE	FERF	CADQ	RRV	ROIL	FNLL	SNAIK	FT	PAG	GRVT : 502						
Cy7822_ZP_03155846	: NGKKL	MDLIND	IL	DYSS	ITS	CDYQ	LN	HEFS	LSY	VNSV	LRD	F	ASEA	QKKS	SID	L	LD	FQV	QKE	END	FYAD	RERIKO	ILSH	L	SNAIK	FT	PD	QGKVT : 531			
Cy8801_ZP_02940946	: SKHLL	EMINE	ILEY	SKLE	AGKY	VLS	IQQF	SLQ	KAAQ	NIHKK	L	RQEA	V	KRKIN	LQEL	QIE	PQEN	FFFAD	PERL	QOILY	HL	LN	NAL	KFT	PE	G	TVI : 531				
CyATCC_YP_001806165	: SKHLL	DMINE	ILEY	SNLQ	SGKY	VLA	REFS	LT	KVAK	NVTQ	RVSD	EAEH	RRIN	LELD	LQIN	Q	Q	S	FYAD	PERV	QOILY	HL	LN	NAL	KFT	PE	NG	SVT : 535			
Ly_ZP_01622671	: SGEHLLE	LINDIL	EL	SEV	AGK	TILQ	SEFSL	TKLAQ	KTLY	SFR	ERAQ	EHNV	LL	ISDI	QV	K	PD	ED	FIAD	QRRV	EKVL	FNLL	SNAIK	FT	PE	SG	QVT : 404				
Mc_YP_002622050	: NGEHL	LALIND	IL	DFSQ	LEAG	AVLN	NEFSL	RYLAQ	Q	TLRS	LKEA	ASSQ	VELI	LD	WQV	PTPK	R	EQAD	PRR	ROIL	FNLL	L	GNAIK	FT	PAG	GQVT : 565					
Mi_YP_001659385	: NGNKL	LELIND	IL	NFSQ	V	SAG	AVLD	LQF	SLKY	CNHV	LEI	IKTEA	ETKE	IN	L	EMK	ITE	KEHY	FYAD	YDRV	QOILY	LYL	K	NAL	KFT	PE	Q	GAVT : 512			
No_NP_485728	: SGEHL	DMIND	IL	DL	SLQIB	AGKTV	LN	ABFSL	VK	TAENTIES	LLEK	AL	SEQVN	LKLD	LQID	PRR	R	FTADA	ARIEQ	ILWN	LL	TNAIK	FT	PE	G	GNVT : 588					
Np_ZP_00108622	: SGEHLLE	MINDIL	DL	SLQIB	AGK	TALN	SEFSL	VN	VAEK	AL	ESLR	NKAT	SE	QIN	LN	LD	LQID	PSR	R	FTADA	ERVA	OILSN	LL	TNAIK	FT	PE	SC	GNVT : 593			
Scy_NP_441517	: NGKKL	LDQINS	II	QL	SOLE	SCQ	TALNC	QSFS	SLHT	LAQTV	HS	LLGVA	IKQ	QIN	LELD	YQIN	V	GDQ	F	CADQ	ERLD	QOIL	TQ	LN	NAL	KFT	PA	E	GTVI : 524		
Syn_YP_001734160	: SCTR	IRNLLQD	IL	DFSQ	TEA	ARS	VLD	LQK	FSLKQ	CYRV	NGFQ	ELG	DRQ	GIN	LK	F	INQ	LE	P	QDSFY	AD	P	IRLEK	ILSH	L	SNAIK	FT	SH	GGEVT : 530		
Th_NP_681689	: SETHLE	ELINS	IL	DL	SEAL	GRS	QLH	RS	AFS	IRQ	CAD	CEV	V	KPQ	AHRQ	VN	LRH	QLM	IP	P	TRDR	F	WGDY	RRIQ	QOIL	IN	LL	SNAIK	FT	PA	MGEVI : 501
Tr_YP_721187	: SGEHLLE	LINDIL	DL	SLQV	EAGK	TVLK	SEFSL	SK	SYEL	QN	FREKA	QNEV	KLIFE	PTV	N	PQD	L	FAAD	QRRV	KOIL	FNLL	S	NGIK	FT	PE	NG	GSVR : 626				

		*	780	*	800	*	820	*	840		*	860	
Sel_Cika	:	LRVW	REGDRAIFQVSDTGIGIPES	EQAQLFQKFQQLDTSIR	ROYGCTGLGLALTKQLVELHGGHIQ	IESTVGQGSTFTVWIPEQT	TLIEPVEPRPSI	:	622				
Ac_YP_001518724	:	LRVW	VEINTVVFQVEDTGIGISS	QLPHLFQKFQQLDGSYHRTYECAGI	GLALAQCCVILHQCWIDV	SSSEGGGSIFTVQLPNQ	SALFQQETSSQP	:	608				
An_YP_321687	:	LRVW	VEDTTSIFQVEDTGIGIP	EEQLPILFEKFQQLDTPYRRRYE	CTGLGLALTKQLVELHRCRIE	VESTVGIGSIFTVWIPYQ	EIRE-----	:	676				
Ar_ZP_03273921	:	LT	VRKDKTAIFQVKDTGIGISKE	QQLPILFEKFQQLDSPYRRQY	GCTGLGLALTKQLVELHGGSIQ	FSEVGLGTKFTVYIPIQ	PLINKSSSNYKT	:	489				
Cr_ZP_00513747	:	LRVW	LKNQLCLEVEDTGIGIEEKI	PILFEFSFQQLENSRRRVY	GCTGLGLALTKQLVELHGGTIE	VESIINKGSTFTVRIPS	QPNKAKLVSQSE	:	627				
Cy0110_ZP_01728511	:	LRIW	RENNQVSLEVEDTGIGIKEE	QIPILFEFSFQQLENSRRRM	YGCTGLGLALTKQLVELHGGTIE	VESIINQGSTFTIVRL	ENQPSQYKSLNNLD	:	630				
Cy7424_ZP_02976063	:	LRIW	REKNQVFQIEDTGIGITQD	LPILFEKFQQLLESSRQRT	HGGTGLGLALTKQLVELHRC	TIEVESTPHEGSLFTV	RLPNSINLKTKRNF	TAD	:	622			
Cy7425_ZP_03139397	:	LS	WVENGKALFQVEDTGIGIS	LSQQLPILFQKFQQLDTSY	TRSYEGVGLGLALTKQLVELH	RCRIEVESTEGIGSIFT	VELPAQSPTATLSA	GLKP	:	598			
Cy7822_ZP_03155846	:	LRIW	REKSQVFQVEDTGIGISQ	EQLPILFEKFQQLDKARQRT	HGGTGLGLALTKQLVELHRC	TIEVESTPEQGSIFTV	RIPNNINLKNKKLL	TNE	:	627			
Cy8801_ZP_02940946	:	LRIW	REGNEAIFQVEDTGIGIAQ	EQLPILFEFSFQQLETSRR	RTYGGTGLGLALTKQLVELH	GCTIEVESTILQEGSIFT	VRLPNQPPRQLKSS	DNLA	:	627			
CyATCC_YP_001806165	:	LRIW	RENNHVSLEVEDTGIGIQ	EKIPLFEFSFQQLENSRRRM	YGCTGLGLALTKQLVELHGG	TIEVESTIINQGSTFI	IRLPNPQNQYKSL	HNLD	:	631			
Ly_ZP_01622671	:	LRV	GVNKNTAIFQIOTDGIGI	SEEQPLFEKFQQLDSSYRR	QYSGTGLGLALTKQLVELH	GCVIKLTSTENVGSTFT	VHIPIQPTPSPNSQ	QVK	:	500			
Mc_YP_002622050	:	LR	WREONRAVFQIEDTGIGI	PQEQPLMFQKFQQLDKTY	QRTYEGTGLGLALTKQLVEL	HGGTIDVESTVGKSI	FTVRVPICSVVPGG	TKSPAS	:	661			
Mi_YP_001659385	:	LRLW	EGSQVIFQVEDTGIGIPA	QKIPILFEKFQQLDGSRR	RYGCVGLGLALTKQLVELH	RCRIEVESTCLDKGSIFT	VRLPQCKPPKPQPS	DNNP	:	608			
No_NP_485728	:	LRIW	VEDTTSIFQVEDTGIGIP	EEQLPILFEKFQQLDTPYRR	RYECTGLGLALTKQLVELH	RCRIEVESTVGIGSIFT	VWIPYQEIRA-----	:	676				
Np_ZP_00108622	:	LRLW	VEDDTAIFQVEDTGIGIR	EEQLPILFEKFQQLDTPYRR	RYECTGLGLALTKQLVELH	RCRIEVESTISIGSIFT	VWIPTQAMRVLS-----	:	679				
Scy_NP_441517	:	LRIW	KSNAIFQVEDTGIGINEQ	QLPVLFQKFVAGDSYTSF	YETGGVGLALTKQLVELH	GCGYIEVESSPGQGTIFT	TVIPQCNFPPTTKG	VQVD	:	620			
Syn_YP_001734160	:	FS	IWRKQKDVQVKDTGIGIP	PQEQPLFEQFQQLDSSMS	RRYDCAGFGLALVKQLVEI	HQGIHVTSTPKKGSIFT	VRIPNQVPRNLHP	VGKGS	:	626			
Th_NP_681689	:	LRA	WKEDELIFQVODTGIGI	PAHLQSLFQKFQQLDSSF	GRAYTCAGLGLALTKQW	VDLHHGWIDVDSTEGK	GSTFTVGLPAISD	PLPDPKPKL	:	597			
Tr_YP_721187	:	LKV	MRYKNTAIFQVEDTGMG	ISQEQPLFEKFQQLNSP	YNRQCGCTGLGLALTKQL	VELHNGIIEVESTVNV	GSTFTVKLPAQQLK	SGNEKRKQK	:	722			

		*	880		*	900	*	920	*	940	*	960	
Sel_Cika	:	DN	-----LPAGH	ILLLEEDEAATVVCEMLTAAG	FKVIWLVDG	STALDQILLQPIVILMAW	PPPDQSCLLILQHLREHQ	ADPHPLVL	:	706			
Ac_YP_001518724	:	LS	-----AGR	IVLIEGHEDATLLCDMLTAAN	YQVIMVVEASAAIDQIRL	LQPIAVIVDAQLPAQGC	LGIIRRLRALPGTDKIKI	IV	:	690			
An_YP_321687	:								:	-			
Ar_ZP_03273921	:	SL	-----PSMYT	DTQPGRIVLIEDDEETATL	ICEILTAAGYQVWVLM	DGLTALATIQLIKEDAFID	LHISGQDGYDIVRHL	REDATTQKIKIVA	:	579			
Cr_ZP_00513747	:	KNQA	-----LFTKN	KNTIVLVESNEEIIATLIG	ELLTAANYHFIWLM	DGTVIKKIELLEPSAVIL	DKELS--KIEI	INDSLKQYPETK	DTKVL	:	713		
Cy0110_ZP_01728511	:	KNRA	-----LFTKN	KNTIILVESNEEIIATLIG	ELLTAANYHFIWLM	DGSKVFKKIELLEPSAVIL	DQDLS--EVLK	INKSLKQLPETK	ETKVL	:	716		
Cy7424_ZP_02976063	:	ENN	-----LINTN	PNNTIILISKDEETATLICE	LLTVNNYQVIWLVD	SYPGIRQIEILHPLIVIL	DQENSQSE--EIV	KALKQFSKTSFI	KVLV	:	709		
Cy7425_ZP_03139397	:	NL	-----PPSDF	LMTGRIVLLESDEETATL	ICDILTAAGYQVWVME	PSTAVEQIQYLQPIAVIT	AVDLPNMDGEDIR	QLRLYPFPPLKILA	:	687			
Cy7822_ZP_03155846	:	TNN	-----VNKSA	QNKTIILISKDEEATLICE	LLTVKNYQVIWLLD	SYPSIRQIEILOPLIVIID	QEI--IQSQEIG	KALKHYPKTSFV	KVLV	:	714		
Cy8801_ZP_02940946	:	INQS	-----LSMGN	RSIVLIESNEELATLIGEL	LTAANYQFIWLM	DSTTAIKKVELFEPTAVIL	DQDLT--DAYKI	SEALKASPKTSIK	VLL	:	713		
CyATCC_YP_001806165	:	KNQA	-----LFTKN	KNTIVLVESNEEIVATLIG	ELLTAANYHFIWLM	DGTVIKKIELLEPSAVIL	DQDLS--EVLEI	NACKELPETQGT	KVLV	:	717		
Ly_ZP_01622671	:	DSAF	KTSLPTTSKNNPLGRF	ILLIENDEEIIATLICE	LLTAVGYQVWVLM	VEGLTALQIQLQPIAVI	VDMLPGQDGYEII	HHLKNTKATQQIKILA	:	596			
Mc_YP_002622050	:	SEK	-----GSSIP	GSIVLIEENNEASATAICE	LLTAAGYHLVWLIES	STAVRQIELLQPHAVIL	DWQLSAMDGYEIS	YIHHKTTTAHIK	VLA	:	748		
Mi_YP_001659385	:	HF	-----NHHHPT	IVLIESDEEIIANLICE	LLMVAHYQVIWLID	SVSAIKKIEIVQPGI	IIVDRKMP--DIYH	CHLLKRSRQTQ	ASKVLI	:	691		
No_NP_485728	:								:	-			
Np_ZP_00108622	:								:	683			
Scy_NP_441517	:	KLDA	-----AMPFN	SSVIVIEQDEEIIATLICE	LLTVANYQVIWLID	TNALQQVELLQPGLI	IIVDGDF--VDV	TEVTRGIKSR	RISKVT	VF	:	706	
Syn_YP_001734160	:	V	-----GGNQGGT	IVLVSQDEEMATLICE	LLTATNYQVIWLID	SEIASRQISALQPI	LIVILDAHQHKIQIED	IDS	LKMAPQTQQIPTLL	:	711		
Th_NP_681689	:	DV	--PPLATTE	VLVEPEGRIVLVSEDEAT	SLICSILTTAGYQVIWL	DGE--VERILALTPIAV	LAEFPSYGDVQEL	VDQLRQ	CTPEQLKIFI	:	689		
Tr_YP_721187	:	GE	-----SLSNV	LSLQGSVVLIEQEEDIATV	ICDILTTAGLKVVWII	EGSTAVEQVILLQPVV	IIDMQLPGINGIEI	IDQLRTTSSSK	NIKFLA	:	812		


		*	980		*	1000	*	
Sel_Cika	:	FLG---	EPPVDPLLTAQASAI	LSKPI	DPQL	LTTH	QGLCPPNLSEGDRPSS	: 754
Ac_YP_001518724	:	LTQEGISKDHQRYLSL	GADAYLL	KS	IQPED	LRKVNVL	-LKSASVL-----	: 735
An_YP_321687	:	-----	-----	-----	-----	-----	-----	: -
Ar_ZP_03273921	:	LT	TN--SDEQEQLQF	GVDKCI	NKPI	ILPNQL	LNQVESLNLMGMTKNYD---	: 625
Cr_ZP_00513747	:	LR	NEITSKEWTEISKMGIDDY	LKPI	IQPNLL	LRVNAL	IFNNNES-----	: 758
Cy0110_ZP_01728511	:	LR	DEITSKEWTEISQMGIDDY	LKPI	IQPNLL	LRVNAL	MFNNNDSENE---	: 764
Cy7424_ZP_02976063	:	LR	KSLKAISWQSLAKKGIDDY	LKPI	IDPTLL	LRKVSFLASIAAHEDKV---	: 757	
Cy7425_ZP_03139397	:	LT	AVQRVPANQLASGFAADAY	LYRP	INPVQL	LDTH	STLFVTPASLS-----	: 733
Cy7822_ZP_03155846	:	LR	TSLQNISWKSILVKNIGIDDY	LKPI	IEPTIL	LRKISFLGSLAIHKDKA---	: 762	
Cy8801_ZP_02940946	:	LS	HQISSTEWTDISKRGIDDY	LKPI	IQPNLL	LRVNAL	MSSDDNEPDDRI-	: 763
CyATCC_YP_001806165	:	LR	DEITSKEWTEISQMGIDDY	LKPI	IQPNLL	LRVNAL	IFSNDKSEDE---	: 765
Ly_ZP_01622671	:	IT	TDSEFTPSES	VF	KTEADDF	ITKPI	QLNQLLKKIMSWNLNSRG-----	: 641
Mc_YP_002622050	:	LL	TSSLSADEQHDLTALVDDY	LPKP	IEPAQL	LLHKVATLMAI	-----	: 789
Mi_YP_001659385	:	LN	DTP-EISVNFLGRHGIDDY	LKPI	IQPSLL	LEKIRYLTAL	-----	: 731
No_NP_485728	:	-----	-----	-----	-----	-----	-----	: -
Np_ZP_00108622	:	-----	-----	-----	-----	-----	-----	: -
Scy_NP_441517	:	LS	ESLSSAEWQALSQKGIDDY	LKPI	QPEL	LTQRVQSI	QQEPLR-----	: 750
Syn_YP_001734160	:	TG	RLSDDQWQKLQKHGFQDY	LPKP	IHSEK	LIDM	NHYVTRHYLASTL---	: 759
Th_NP_681689	:	LG	SKGNYQ-----	GVDRY	IPLP	IHPESFL	QQVTMGLTSLATSAQ----	: 728
Tr_YP_721187	:	LT	TLNTEINREYCDAIGVDEC	ITKPV	NLEY	LNKMIHL	-LAN-----	: 853

Figure S3

